Stochastic Models in Population Genetics:
Genealogy and Genetic Differentiation
in Structured Populations

by Hilde Maria Jozefa Dominiek Herbots

Thesis submitted for the degree of
Doctor of Philosophy (PhD)
of the University of London

Queen Mary and Westfield College
September 1994
ABSTRACT

The theory of probability and stochastic processes is applied to a current issue in population genetics, namely that of genealogy and genetic differentiation in subdivided populations.

It is proved that under a reasonable model for reproduction and migration, the ancestral process of a sample from a subdivided population converges weakly, as the subpopulation sizes tend to infinity, to a continuous-time Markov chain called the "structured coalescent".

The moment-generating function, the mean and the second moment of the time since the most recent common ancestor (called the "coalescence time") of a pair of genes are calculated explicitly for a range of models of population structure.

The value of Wright's coefficient $F_{ST}$, which serves as a measure of the subpopulation differentiation and which can be related to the coalescence times of pairs of genes sampled within or among subpopulations, is calculated explicitly for various models of population structure. It is shown that the dependence of $F_{ST}$ on the mutation rate may be more marked than is generally believed, particularly when gene flow is restricted to an essentially one-dimensional habitat with a large number of subpopulations.

Several more general results about genealogy and subpopulation differentiation are proved. Simple relationships are found between moments of within and between population coalescence times. Weighting each subpopulation by its relative size, the asymptotic behaviour of $F_{ST}$ at large mutation rates is independent of the details of population structure. Two sets of symmetry conditions on the population structure are found for which the mean coalescence time of a pair of genes from a single subpopulation is independent of the migration rate and equal to that of two individuals from a panmictic population of the same total size. Under graph-theoretic conditions on the population structure, there is a uniform relationship between the $F_{ST}$ value of a pair of neighbouring subpopulations, in the limit of zero mutation rate, and the migration rate.
Soli Deo Honor et Gloria

Queen Mary College
ACKNOWLEDGEMENTS

My first words of gratitude go to Peter Donnelly for being a wonderful supervisor. Without his invaluable advice and encouragement this thesis would never have been written. I sincerely thank him for all his help.

It is a pleasure to acknowledge my debt to Jef Teugels who, knowing of my interest in mathematical genetics, made it possible for me to meet Peter and assisted in obtaining financial support for my studies. I have always been able to rely on his encouragement and advice.

I thank the members of the Mathematical Genetics Group and the Probability and Statistics Group at Queen Mary and Westfield College for making me feel very much part of a lively and friendly research community. I particularly wish to thank David Balding, Mark Beaumont, Charles Goldie, Richard Nichols and Eliane Rodrigues.

I am grateful also to Herman Van Den Berghe for his interest and encouragement, and to Monty Slatkin and Walter Van Assche for helpful suggestions.

The author is a Research Assistant of the Belgian National Fund for Scientific Research. I also thank the Science and Engineering Research Council and the Drapers' Company for their financial support.

I am specially grateful to my family and to William for their loving support and encouragement throughout my studies. Many thanks also to my friends in Ashwell House for providing such a homely and happy environment in which to live and work all these years.
To my Parents
Contents

1 INTRODUCTION .............................................. 11
  1.1 Kingman’s coalescent ..................................... 13
  1.2 The structured coalescent .................................. 14

2 THE STRUCTURED COALESCENT ................................. 19
  2.1 The discrete-time model .................................... 20
  2.2 Convergence to the structured coalescent ......................... 22
  2.3 Coalescence times ......................................... 33
    2.3.1 The coalescence time of a sample from the population ....... 34
    2.3.2 The “k to k − 1” coalescence time of a sample from the population .... 38

3 GENEALOGY UNDER VARIOUS MODELS OF POPULATION STRUCTURE .... 40
  3.1 The symmetric island model .................................. 41
  3.2 The stepping-stone models ................................... 47
    3.2.1 The circular stepping-stone model ......................... 48
    3.2.2 The infinite linear stepping-stone model .................... 50
    3.2.3 Two-dimensional stepping-stone model on a torus ............ 52
    3.2.4 The infinite two-dimensional stepping-stone model .......... 57
  3.3 Some less symmetric structures ................................ 61
    3.3.1 The general two-population model .......................... 61
    3.3.2 The continental island model .............................. 63
    3.3.3 Four-population model with geographic barrier ............ 67

4 SUBPOPULATION DIFFERENTIATION UNDER VARIOUS MODELS OF POPULATION STRUCTURE .... 70
  4.1 Wright’s coefficient $F_{ST}$ ................................. 71
    4.1.1 Introduction ........................................... 71
    4.1.2 $F_{ST}$ in terms of coalescence times ....................... 73
4.2 The symmetric case ........................................... 75
4.2.1 The symmetric island model ............................... 76
4.2.2 The circular stepping-stone model .......................... 81
4.2.3 The infinite linear stepping-stone model ......................... 87
4.2.4 Two-dimensional stepping-stone model on a torus .................. 90
4.2.5 The infinite two-dimensional stepping-stone model ................. 95
4.2.6 The dependence of $F_{ST}$ on the mutation rate .................. 99
4.2.7 Discussion .................................................. 105
4.3 Some less symmetric structures ................................. 108
4.3.1 The general two-population model ......................... 110
4.3.2 The continental island model ................................ 117
4.3.3 Four-population model with geographic barrier ............... 124
4.3.4 Discussion .................................................. 128

5 GENERAL RESULTS ............................................. 131
5.1 Genealogical results under weighting by size. .................. 132
5.2 The asymptotic behaviour of $F_{ST}$ at large mutation rates. ... 134
5.3 The mean coalescence time of two individuals from a single subpopulation. .... 136
5.4 The $F_{ST}$ value of a pair of neighbouring subpopulations. ... 148

REFERENCES ..................................................... 152
List of Figures

Figure 3.1 .............................................. 41
Figure 3.2 .............................................. 44
Figure 3.3 .............................................. 48
Figure 3.4 .............................................. 51
Figure 3.5 .............................................. 57
Figure 3.6 .............................................. 61
Figure 3.7 .............................................. 64
Figure 3.8 .............................................. 67

Figure 4.1 .............................................. 78
Figure 4.2 .............................................. 80
Figure 4.3 .............................................. 81
Figure 4.4 .............................................. 83
Figure 4.5 .............................................. 85
Figure 4.6 .............................................. 86
Figure 4.7 .............................................. 88
Figure 4.8 .............................................. 89
Figure 4.9 .............................................. 91
Figure 4.10 ............................................. 92
Figure 4.11 ............................................. 93
Figure 4.12 ............................................. 94
Figure 4.13 ............................................. 97
Figure 4.14 ............................................. 98
Figure 4.15 ............................................ 100
Figure 4.16 ............................................ 102
Figure 4.17 ............................................ 104
Figure 4.18 ............................................ 112
Figure 4.19 ............................................ 113
NOTATION

Multi-dimensional objects and matrices are printed in bold. In addition, the following notations are used.

$\equiv$ : equality by definition.

$\Delta$ : equality in distribution.

$\Leftrightarrow$ : if and only if.

$\mathbb{N}$ : the set of the natural numbers, including zero: $\mathbb{N} := \{0, 1, 2, \ldots\}$.

$\mathbb{N}_0$ : the set of the natural numbers, not including zero: $\mathbb{N}_0 := \mathbb{N} \setminus \{0\}$.

$\mathbb{Z}$ : the set of the integers.

$\mathbb{Z}_q$ : the set of the integers modulo $q$.

$\mathbb{R}$ : the set of the real numbers.

$\mathbb{N}^S$ : the set of all functions $g : S \to \mathbb{N}$.

$\mathcal{L}_1$ : the normed vector space of the integrable real-valued random variables (that is, real-valued random variables $X$ with $E|X| < \infty$).

$\delta_{ij}$ : the Kronecker delta: $\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$

$[x]$ : the largest integer less than or equal to $x$.

$|\alpha|$ : $|\alpha| := \sum_{i \in S} \alpha_i$, if $\alpha \in \mathbb{N}^S$.

$|\xi|$ : $|\xi|$ := the number of equivalence classes in $\xi$, if $\xi$ is an equivalence relation.

$\sim$ : $[f(x) \sim g(x) \text{ as } x \to \infty] \Leftrightarrow [\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1]$. 
Chapter 1

INTRODUCTION

In stochastic modelling, the powerful mathematical framework of probability and stochastic processes is applied to describe, simulate and predict real world phenomena whose behaviour involves some randomness. One growing and fascinating area of stochastic modelling is in genetics. In the application of probability theory to genetics, this thesis touches two scientific disciplines which are both relatively young and which are both going through an exciting period of progress.

Population genetics is concerned with the genetic composition and evolution of populations. The genetics of a population is intrinsically related to its genealogy, and the last two decades have seen the development of numerous results on the genealogy of populations whose genetic composition is described by the classical (idealized) population genetics models. For many of these models, the intimate relationship between genealogy and genetic composition is now well understood. By bringing the genealogy into the open, a deeper insight is gained into the genetic and evolutionary mechanisms at work, while a firm basis is provided to address current questions concerning genetic variation.

A major development in the study of genealogy has been the so-called “coalescent approach”, which models genealogy backward in time. The basic idea of the coalescent approach is to draw a sample from the population and to trace back the ancestry of the individuals in the sample, focusing on the times in the past when two or more individuals in the sample have a common ancestor. One advantage of this approach backward in time is that only those individuals in the past who have descendants in the current sample need to be considered, which considerably reduces the complexity of simulations. Characteristic of the coalescent approach is also the “diffusion timescale” approximation: rather than measuring time in discrete generations, each time-unit is made up of a number of generations which is typically of the order of the population size, and one takes the limit as the population size tends to infinity. The stochastic processes arising in this context are time-homogeneous Markov chains in continuous time, which lend themselves very well to analytical study and which make it relatively easy to obtain explicit results.
Although the idea of tracing ancestral lineages was not new in genetics (for example, the related concept of "identity by descent" was introduced by Malécot 1948, 1969, 1975, while Watterson 1975 already described the basic properties of genealogies under neutral mutation models), it was Kingman (1980, 1982a,b,c) who formalized the coalescent approach and provided the underlying mathematical theory, which has received considerable attention since and which is still being extended. Similar genealogical ideas were introduced independently, at about the same time as Kingman's work, by Griffiths (1980, 1981), Hudson (1983a,b) and Tajima (1983).

Kingman proved that for a wide range of reproductive models, the family relationships in a sample from the population are well described by a continuous-time Markov chain called the "coalescent". Starting from a sample from the population at time zero and tracing back their ancestry, the coalescent partitions the sample into equivalence classes, individuals belonging to the same class at time $t \geq 0$ if they had the same ancestor time $t$ ago. As $t$ increases and we look further back into the past, the number of distinct ancestors of the sample decreases and the equivalence classes coalesce, until ultimately the whole sample forms one equivalence class sharing a single common ancestor.

Whereas Kingman's coalescent assumes an unstructured population, most real populations do have geographic or other structure. This thesis is concerned with populations that are subdivided into distinct subpopulations which are partially isolated from each other. The family relationships in such a population are, under reasonable assumptions about reproduction and migration, well described by a stochastic process (similar to Kingman's coalescent) which is called the "structured coalescent".

In this introduction we set out the fundamental assumptions of the coalescent approach and we briefly introduce Kingman's coalescent (Section 1). In Section 2 we describe the structured coalescent, the stochastic process on which this thesis is based, and we give an overview of the chapters to follow.

The coalescent approach has been used in a wide variety of applications, too large to be exhausted here. See Tavaré (1984) and Hudson (1990) for review.

For the purposes of this thesis, a gene is an indivisible string of genetic material which can be transmitted from parent to offspring. Within each cell of an organism, genes are arranged linearly along chromosomes, which are, essentially, very long strings of genetic material which typically contain thousands of genes. The position of a gene along a chromosome is called a locus. Whenever we write about genetics in this work, we are concerned with genes at a single locus. Within a population, genes at a particular locus can be of different types; these alternative types are called alleles. Genes can change from one allele to another by mutation.
Chapter 1: INTRODUCTION

1.1 Kingman's coalescent

All coalescent theory relies on two key assumptions. The first is that individuals are haploid, which means that each individual has only one parent. While this is true for some species, most populations of interest (for example, the human population) are diploid, that is, each individual has two parents. However, the coalescent approach can apply to diploid populations if, rather than working with the diploid individuals, one works with the genes at a particular locus, which are haploid. The second assumption is that the individuals are neutral, that is, there are no inherited differences of fitness or fertility between individuals. In the case of genes, this assumption means that there is no selection affecting the locus in question.

Kingman's coalescent (Kingman 1982a,b,c) assumes in addition that the population is panmictic. For a haploid population this means that, fixing an individual in a particular generation, all individuals in the previous generation are equally likely to have been its parent, i.e. the population has no substructure or subdivision. Kingman's work also assumes that the population size is constant in time. However, as was pointed out by Kingman (1982a), the latter assumption is not essential and the coalescent applies in many contexts to populations of variable size if one chooses an appropriate non-linear time-scale (see also Marjoram and Donnelly 1994b).

The mathematical form in which the above four assumptions are used is by assuming an "exchangeable" model of reproduction (introduced by Cannings 1974), in which the offspring numbers of the individuals in each generation are exchangeable random variables, offspring numbers in different generations being independent. The most commonly assumed exchangeable model is the neutral Wright-Fisher model, in which the offspring numbers have a symmetric multinomial distribution (see Cannings 1974 or Ewens 1979, or see Section 2.1). Another frequently used exchangeable model is the Moran model (see Moran 1958 or Cannings 1974).

Consider a population of constant size $N$ individuals which evolves in discrete generations. At a particular generation (which we call time zero) a sample of fixed size $n_0$ individuals is drawn from the population, and we label the individuals in the sample as $1, 2, \ldots, n_0$. Tracing the ancestral lineages of the individuals in the sample, Kingman (1982a,b,c) defined a genealogical process $\mathcal{R}_N \equiv \{\mathcal{R}_N(\tau) : \tau = 0, 1, 2, \ldots\}$ whose state space is the set $\mathcal{E}_{n_0}$ of all equivalence relations on $\{1, 2, \ldots, n_0\}$ and where $(i, j) \in \mathcal{R}_N(\tau)$ if and only if individuals $i$ and $j$ have the same ancestor $\tau$ generations ago. Clearly $\mathcal{R}_N(0) = \{(i, i); i = 1, 2, \ldots, n_0\}$ and $\mathcal{R}_N(\tau_1) \subseteq \mathcal{R}_N(\tau_2)$ if $\tau_1 \leq \tau_2$, since two individuals sharing the same ancestor $\tau_1$ generations ago will also have the same ancestor $\tau_2$ generations ago. Each equivalence class of $\mathcal{R}_N(\tau)$ corresponds to a distinct ancestor of the sample alive $\tau$ generations ago. If two such ancestors have the same parent in the previous generation ($\tau + 1$ generations ago), the two equivalence classes of $\mathcal{R}_N(\tau)$ that correspond to these ancestors merge, or "coalesce", to form one equivalence class in $\mathcal{R}_N(\tau+1)$. If individuals choose their parent at random, independently and uniformly from among the individuals of the previous generation
(which is the backward structure of the neutral Wright-Fisher model), the probability that two distinct ancestors \( \tau \) generations ago have the same parent (and their equivalence classes coalesce) is \( 1/N \). We denote the number of equivalence classes of an equivalence relation \( \xi \in \mathcal{E}_{n_0} \) by \( |\xi| \). For \( \xi, \eta \in \mathcal{E}_{n_0} \), we write \( \xi \prec \eta \) if \( \eta \) can be obtained from \( \xi \) by combining two equivalence classes of \( \xi \) into a single equivalence class of \( \eta \), i.e.

\[ \xi \prec \eta \Leftrightarrow \xi \subset \eta \text{ and } |\eta| = |\xi| - 1. \]

Assuming the neutral Wright-Fisher model of reproduction and measuring time in units of \( N \) generations, Kingman (1982a) proved that the process \( \mathcal{R}_N \) converges weakly, as \( N \) tends to infinity, to the coalescent, which is the continuous-time Markov chain \( \{\mathcal{R}(t) : t \geq 0\} \) whose Q-matrix \( \mathcal{R} = (R_{\xi,\eta}) \) is given by

\[ R_{\xi,\eta} = \begin{cases} \frac{-|\xi|}{2} & \text{if } \eta = \xi \\ 1 & \text{if } \xi \prec \eta \\ 0 & \text{otherwise} \end{cases} \tag{1.1} \]

for \( \xi, \eta \in \mathcal{E}_{n_0} \). Kingman (1982c) proved that this result is robust, in the sense that it is valid for the entire class of the exchangeable models satisfying certain (weak) conditions on the moments of the distribution of the offspring number of a typical individual, provided an appropriate time-scaling is used (usually in units of \( N/\sigma^2 \) generations, where \( \sigma^2 \) is the variance of the offspring number of a typical individual, in the limit as \( N \to \infty \)).

In many applications it is irrelevant which individuals are related to each other and it suffices to keep account of the number of distinct ancestors of the sample at each time in the past. A simpler version of the coalescent is the ancestral process \( \{ \alpha(t) : t \geq 0 \} \), where \( \alpha(t) := |\mathcal{R}(t)| \) gives the number of distinct ancestors of the sample, time \( t \) ago. This ancestral process is a pure death process whose transition rate from \( k \) to \( k - 1 \) is \( \binom{k}{2} \), as \( k \) ancestral lineages give \( \binom{k}{2} \) pairs of lineages which can coalesce and every two lineages coalesce at rate 1 (see (1.1)). In particular, the time since \( k \) individuals descended from \( k - 1 \) distinct ancestors is exponentially distributed with mean \( \binom{k}{2}^{-1} \).

### 1.2 The structured coalescent

Whereas Kingman's coalescent assumes a panmictic population, most real populations are geographically structured. As a model of a structured population we will consider in this thesis populations that are subdivided into large colonies of fixed size which interchange genes by migration. A detailed such model will be given in Chapter 2. When modelling genealogy in a subdivided population, it is not sufficient to trace the ancestry of the individuals in the sample under study, but one also needs to keep track of the locations of the ancestors of the sample at each time in the
past, if the Markov character of the coalescent is to be maintained. The resulting coalescent-like process for subdivided populations is called the "structured coalescent". The ancestral process associated with the structured coalescent was introduced by Takahata (1988) for a population consisting of only two subpopulations, and formulated in its general form by Notohara (1990). For simplicity, we refer to this ancestral process as the structured coalescent.

The population considered is haploid and divided into subpopulations or "colonies" which are all large and panmictic and which are partially isolated from each other. We denote by $S$ the set of the subpopulation labels ($S$ may be finite or infinite). The size of subpopulation $i$ is $N_i = 2c_iN$ haploid individuals, where $c_i$ is a positive integer constant and $N$ is large. In diploid applications, $2c_iN$ is the number of genes in subpopulation $i$, at a particular locus, so that $c_iN$ is the number of diploid individuals in subpopulation $i$. At a particular generation which we call time zero, we draw a sample of fixed size $n_0$ individuals from the total population (where $n_0$ is finite), and we trace the ancestry of the individuals in the sample. At each time in the past, we count how many distinct ancestors the $n_0$ sampled individuals have in each subpopulation. We denote by $a(r)$ the number of distinct ancestors the sample has in subpopulation $i$, $r$ generations ago ($i \in S$), and by $\alpha_N(r)$ the ordered set $\{a(r)_{i \in S}\}$, with a component for each subpopulation. If there are $K$ subpopulations and $K$ is finite, then $\alpha_N(r)$ is a $K$-tuple. If the number of subpopulations is infinite, $\alpha_N(r)$ is a sequence with index set $S$. In standard mathematical notation, we write that $Q_N(T) \in \mathcal{N}^S$, which is the set of all functions from $S$ to $\mathbb{N}$, where $\mathbb{N}$ is the set of the natural numbers, including zero: $\mathbb{N} = \{0, 1, 2, \ldots\}$. The process $\alpha_N = \{\alpha_N(r) : r = 0, 1, 2, \ldots\}$ will be called the ancestral process.

Tracing the ancestral lineages of the individuals in the sample, two types of events can occur. Two particular lineages can coalesce at the most recent common ancestor of the corresponding individuals in the sample (this can only occur when these lineages reside in the same subpopulation), in which case the number of distinct ancestors in that subpopulation (i.e. the value of $\alpha_N(r)$) decreases by one. The rate at which such a coalescence event occurs is, for many exchangeable models of reproduction (Cannings 1974), inversely proportional to the size of the subpopulation. If an ancestor in subpopulation $i$ is an immigrant from subpopulation $j$ (which we also describe as a "migration" of the ancestor from subpopulation $i$ to subpopulation $j$ backward in time), the number of distinct ancestors in subpopulation $i$ decreases by one, while that in subpopulation $j$ increases by one. Following the notation of Notohara (1990), we denote by $e^i$ the element of $\mathcal{N}^S$ with components

$$(e^i)_j = \delta_{ij} = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

and we define addition and subtraction of elements of $\mathcal{N}^S$ to be component-wise, i.e. the sum or difference of two sequences (or $K$-tuples) in $\mathcal{N}^S$ is obtained by adding or subtracting their
corresponding components. If $\alpha_N(\tau) = \alpha$ and two lineages in subpopulation $i$ coalesce, the value of $\alpha_N(\tau)$ is changed to $\alpha - \epsilon^i$; the migration of an ancestral lineage from subpopulation $i$ to subpopulation $j$ (backward in time) changes the value of $\alpha_N(\tau)$ from $\alpha$ to $\alpha - \epsilon^i + \epsilon^j$.

Under reasonable assumptions about reproduction and migration, the ancestral process $\alpha_N$ is, with the appropriate re-scaling of time, well approximated by the "structured coalescent", which is the continuous-time Markov chain $\{\alpha(t); t \geq 0\}$ with Q-matrix $Q$ whose entries are

$$Q_{\alpha, \beta} = \begin{cases} \sum_{i \in S} \left( \alpha_i \frac{M_{ii}}{2} + \frac{1}{\epsilon_i} \left( \alpha_i \frac{M_{ii}}{2} \right) \right) & \text{if } \beta = \alpha \\ \frac{\alpha_i M_{ij}}{2} \left( \frac{\alpha_j}{2} \right) & \text{if } \beta = \alpha - \epsilon^i + \epsilon^j \ (j \neq i) \\ 0 & \text{if } \beta = \alpha - \epsilon^i \end{cases}$$

(Notohara 1990), where $M_{ij}$ is the scaled migration rate (forward in time) from subpopulation $j$ to subpopulation $i$ and $M_i = \sum_{j \neq i} M_{ij}$. For example, if reproduction in each subpopulation follows the neutral Wright-Fisher model and a constant proportion $q_{ji}$ of the individuals born in subpopulation $j$ migrate to subpopulation $i$ every generation ($i, j \in S$), time-scaling is in units of $2N$ generations and $M_{ij} = \lim_{N \to \infty} \left( 4N \frac{q_{ji}}{\epsilon_i q_{ij}} \right)$. Intuitively the matrix $Q$ is understood as follows: tracing the ancestral lineages of a sample from the population, any two lineages in subpopulation $i$ have coalescence rate $1/\epsilon_i$, while the rate at which a lineage moves from subpopulation $i$ to subpopulation $j$ is $M_{ij}/2$ (the factor of $1/2$ is standard and is convenient in applications involving pairs of genes).

The structured coalescent has proved valuable in understanding and modelling the genetic and demographic processes of interest in a variety of applications concerning structured populations. Applications of the coalescent approach to subdivided populations include those by Griffiths (1981), Takahata (1988, 1991), Tajima (1989), Notohara (1990), Hudson (1990) and references in the latter. Marjoram and Donnelly (1994a,b) used the coalescent approach to simulate structured populations of variable size, in a study of pairwise comparisons of mitochondrial DNA sequences and the time since Eve, the most recent common ancestor of human mitochondrial DNA.

Studying Wright's coefficient $F_{ST}$ (Wright 1951), which serves as a measure of the genetic differentiation among subpopulations and which is used to estimate the effective level of gene flow between subpopulations, Slatkin (1991) related $F_{ST}$ in an approximate way to the mean coalescence times of pairs of genes drawn within and between subpopulations. Slatkin (1993) showed these results provide a way of testing for isolation by distance in a natural population. In fact, $F_{ST}$ can be expressed in an exact way in terms of the moment-generating functions of coalescence times of pairs of genes. This will be done in Chapter 4, where we will also examine the dependence of $F_{ST}$ on the parameters of population structure and on the mutation rate.

This thesis is concerned with the structured coalescent. In Chapter 2 a rigorous proof is
given that, under a particular discrete-time model for reproduction and migration, the genealogy of a sample from a subdivided population is indeed, in the diffusion time-scale approximation, described by the structured coalescent. This result was stated by Notohara (1990) for a different model for migration and reproduction. However, certain steps in Notohara's proof are not entirely clear to us (more details will be given in Chapter 2). By standard probabilistic methods we derive systems of linear equations for the mean, the second moment and the moment-generating function of the time since the most recent common ancestor of a sample from the population, and for those of the time since $k$ sampled individuals descended from $k - 1$ distinct ancestors.

In Chapter 3, these equations are solved for a sample of two individuals under a range of models of population structure, leading to explicit formulae for the mean, the second moment and the moment-generating function of the time since the most recent common ancestor of a pair of individuals under the symmetric island model, the finite and infinite stepping-stone models in one and two dimensions and some less symmetric structures (a general two-population model, a "continental island model" and a four-population model with geographic barrier). For the symmetric island model, a more complete analysis is given and the effect of the population subdivision on the coalescence time of a pair of individuals is shown explicitly by comparison with a panmictic population.

Chapter 4 is concerned with the genetic differentiation among subpopulations of a subdivided population. Wright's coefficient $F_{ST}$, which serves as a measure of the amount of subpopulation differentiation, is expressed in exact and approximate ways in terms of the time since the most recent common ancestor of a pair of genes sampled from a single subpopulation and the time since the most recent common ancestor of two genes sampled from the total population. Using the genealogical results of Chapter 3, the values of $F_{ST}$ and its approximations are calculated explicitly for the various models of population structure introduced in Chapter 3. The dependence of $F_{ST}$ on the parameters of population structure and, in particular, on the neutral mutation rate is studied. The accuracy of Slatkin's approximation for $F_{ST}$ (Slatkin 1991) is also examined.

Chapter 5 contains a variety of "general" results on both genealogy and subpopulation differentiation. Some simple relationships are given (valid when the different subpopulations are weighted by their relative sizes) between the moments of the coalescence time of two individuals from a single subpopulation and those of the coalescence time of two individuals sampled from the total population, and between the corresponding probabilities of identity by descent. We prove that under weighting by size, the asymptotic behaviour of $F_{ST}$ for very large mutation rates is independent of the details of population structure and dependent only on the mutation rate and on the number of subpopulations considered. The link between population structure and graph theory is made and two sets of (graph-theoretic) symmetry and regularity conditions on the population structure are given under which the mean time since the most recent common ancestor of a pair of
individuals from any single subpopulation is independent of the migration rate and equal to that of two individuals from a panmictic population of the same total size. We also show that, under graph-theoretic symmetry and regularity conditions, there is a uniform relationship between the $F_{ST}$ value of a pair of neighbouring subpopulations, in the limit of zero mutation rate, and the migration rate.

Note that the work in this thesis is exact only for haploid species, and for diploid species with exclusively gametic migration (that is, migration of the sex cells, which are haploid; gametic migration occurs in many species of plants). Nagylaki (1983) and Sawyer (1976) have set out conditions under which models of truly diploid migration are well approximated by the model of gamete migration. The results on probability of identity by descent and on subpopulation differentiation are for selectively neutral genes at a single locus, subject to mutation and migration but without intragenic recombination. The results on $F_{ST}$ are exact only for the infinite alleles model of mutation, in which every mutation produces a novel allele.
Chapter 2

THE STRUCTURED COALESCENT

The major goal of this chapter is to show that under reasonable assumptions about reproduction and migration, genealogy in a subdivided population is well described by the structured coalescent.

Considering a population divided into a finite or infinite number of subpopulations and evolving in discrete generations, Notohara (1990) showed that the genealogy of a sample from this population is approximately described by the structured coalescent. In Notohara's model, each generation is made up of two discrete steps, the first one due to migration and the second one due to reproduction. In the migration step the individuals migrate independently between the subpopulations. After the migration step, Wright-Fisher type reproduction brings the size of each subpopulation back to its size before migration. Although under this model the individuals migrate independently forward in time, their migration is not independent backward in time, as migration backward in time necessarily brings the size of each subpopulation to a fixed number, constant in time. Therefore, it is not entirely clear to us that Notohara's argument, which inappropriately relies on individuals migrating independently backward in time, is valid. Some minor additional complications may arise from the fact that under Notohara's model, the mechanisms of migration and reproduction are not independent, since at reproduction, the offspring distribution of the individuals in a particular subpopulation depends on the number of individuals present in that subpopulation after the migration step, from the fact that the offspring distributions (and hence coalescence events when we look backward in time) in different subpopulations are not independent, as the subpopulation sizes after migration are not independent, and from the possibility that migration empties a subpopulation. It is plausible (but not obvious) that these problems disappear in the limit as the subpopulations become infinitely large. In this chapter (Section 2) it is proved in detail that, for a slightly simpler model than Notohara's, the ancestral process
converges to the structured coalescent.

The first section of this chapter contains a precise description of the population model we will assume throughout this thesis. In Section 2 it is proved that under this model, the ancestral process converges weakly to the structured coalescent as the subpopulations become large. In Section 3, a first-step argument yields systems of linear equations for the moment-generating function, the mean and the second moment of, respectively, the time since the most recent common ancestor of a sample from the population (Subsection 3.1) and the time since a sample of \( k \) individuals descended from \( k - 1 \) distinct ancestors (Subsection 3.2). These systems of equations will be solved in Chapter 3 for samples of size two under various models of population structure.

### 2.1 The discrete-time model

We consider a haploid population divided into a finite or infinite number of subpopulations which are all large and panmictic and which are partially isolated from each other. We denote by \( S \) the set of the subpopulation labels, where we assume that \( S \) is countable. The subpopulation sizes, migration pattern and migration rates are assumed to be constant in time. The number of haploid individuals in subpopulation \( i \) is \( N_i = 2c_iN \), where \( c_i \) is a positive integer constant and \( N \) is large.

(In diploid applications, \( 2c_N \) is the number of genes in subpopulation \( i \), at a particular locus, rather than the number of individuals.) Denoting \( c = \sum_{i=1}^{n} c_i \), the total population size is \( 2cN \) (where \( c = \infty \) if the number of subpopulations is infinite). The population evolves in discrete non-overlapping generations. Every generation is made up of two discrete steps, the first one due to reproduction, the second one due to migration.

Reproduction is haploid (each member of a certain generation is the child of exactly one member of the previous generation) and takes place within each subpopulation according to the neutral Wright-Fisher model. This means that in each subpopulation, the joint distribution of the offspring numbers of the different individuals at any particular generation is symmetric multinomial, maintaining the size of that subpopulation: denoting by \( Y_{ij} \) the number of offspring of the \( j \)th individual in subpopulation \( i \) \((i \in S, j = 1, \ldots, N_i)\),

\[
P \{ Y_{i1} = y_1, \ldots, Y_{iN_i} = y_{N_i} \} = \begin{cases} \frac{N_i!}{y_1! \ldots y_{N_i}!} \left( 1 - \frac{y_1}{N_i} \right)^{N_i - y_1} & \text{if } \sum_{j=1}^{N_i} y_j = N_i \\ 0 & \text{otherwise.} \end{cases}
\]

Offspring numbers in different subpopulations or at different generations are independent. This description forward in time is equivalent to the following backward description: each individual in subpopulation \( i \) immediately after the reproduction step chooses its parent at random, independently and uniformly from among the \( N_i \) individuals in subpopulation \( i \) just before reproduction. This simple structure backward in time is the major advantage of the Wright-Fisher model for reproduction. While the convergence result proved in the next section explicitly assumes the neutral
Wright-Fisher model, it is likely to be valid (under mild moment conditions) for any exchangeable model of reproduction (Cannings 1974), provided an appropriate time-scale is used; if in the limit as $N \to \infty$, the variance of the number of offspring of an individual in a subpopulation varies among subpopulations, the coalescence rates $1/c_i$ in (2.2) also need to be adapted.

At each generation, the reproduction step is followed by a migration step. We assume that in every generation, a fixed proportion $q_{ij}$ ($q_{ij} \geq 0$, $\sum_{j \neq i} q_{ij} \leq 1$) of the individuals born in subpopulation $i$ migrate to subpopulation $j$ (these migrants are chosen at random, independently and uniformly, without replacement, from subpopulation $i$), where $q_{ij}$ is constant in time $(i, j \in \mathcal{S})$. It is assumed that the size of each subpopulation is maintained under migration, which requires

\begin{equation}
\forall i \in \mathcal{S} : c_i \sum_{j \neq i} q_{ij} = \sum_{j \neq 1} c_j q_{ji}.
\end{equation}

While each individual in subpopulation $i$ has probability $q_{ij}$ of migrating to subpopulation $j$, the requirement of constant subpopulation sizes implies that the individuals do not migrate independently of each other. Among the $N_i$ individuals making up subpopulation $i$ just after the migration step at any particular generation, there are $N_j q_{ji}$ who are immigrants from subpopulation $j$. The "backward migration rate" $m_{ij}$ from subpopulation $i$ to subpopulation $j$, defined as the proportion of the individuals in subpopulation $i$ immediately after the migration step who were born in subpopulation $j$, is therefore given by

$$m_{ij} = \frac{N_j q_{ji}}{N_i} = \frac{c_j}{c_i} q_{ji} \quad (i, j \in \mathcal{S}).$$

We also denote $m_i = \sum_{j \neq i} m_{ij}$, which is the proportion of the individuals in subpopulation $i$ after migration who were born in another subpopulation.

Assume that the population has been evolving in this way indefinitely. At a particular generation which we call time zero, we draw a sample of $n_0$ individuals from the total population (where $n_0$ is finite and fixed) and we trace back the ancestry of the individuals in the sample, generation by generation. At each time in the past we count how many distinct ancestors the $n_0$ sampled individuals have in each subpopulation. We define the ancestral process $\alpha_N = \{\alpha_N(\tau); \tau = 0, 1, 2, \ldots\}$, where $\forall \tau \in \mathbb{N}$ : $\alpha_N(\tau) \in \mathbb{N}^\mathcal{S}$ with $i$th component $\alpha_N^i(\tau)$ denoting the number of distinct ancestors the sample has in subpopulation $i$, $\tau$ generations ago $(i \in \mathcal{S})$. As an initial sample of fixed size $n_0$ can have at most $n_0$ distinct ancestors at any time, the state space of this ancestral process is the set

$$E := \{\alpha \in \mathbb{N}^\mathcal{S} : \sum_{i \in \mathcal{S}} \alpha_i \leq n_0\}.$$

This set is infinite when the number of subpopulations is infinite. Each element $\alpha$ of $E$ can be obtained by assigning $\alpha = \sum_{i \in \mathcal{S}} \alpha_i$ individuals to the subpopulations in $\mathcal{S}$ and counting the number of individuals thus assigned to each subpopulation. As different elements of $E$ require different assignments of individuals, $E$ is in one-to-one correspondence to a subset of $\bigcup_{a=0}^{n_0} \mathcal{S}^a$. As
Chapter 2: THE STRUCTURED COALESCENT

$S$ is countable, $\bigcup_{s=0}^{\infty} S^s$ is countable and hence $E$ is countable. We denote by $D_E[0, \infty)$ the space of right-continuous functions from $[0, \infty)$ into $E$ having limits from the left. It is proved below (see the next section) that, measuring time in units of $2N$ generations and making the appropriate assumptions, the ancestral process converges weakly to the structured coalescent, as $N$ tends to infinity.

2.2 Convergence to the structured coalescent

All of the work presented in this thesis is based upon the assumption that the structured coalescent is an appropriate description of genealogy. We now proceed to establish that, under the model set out in the previous section, genealogy is indeed well described by the structured coalescent.

Theorem 2.1 Assume that

(i) $\forall i, j \in S$ with $j \neq i$, $4N_{m_{ij}}$ increases monotonically with increasing $N$, with

$$\lim_{N \to \infty} (4N_{m_{ij}}) = M_{ij},$$

and, denoting $M_i := \sum_{j \neq i} M_{ij}$, assume that

(ii) $\sup_{i \in S} M_i < \infty$.

Then as $N \to \infty$, the process $\{\alpha_N([2Nt]) : t \geq 0\}$ converges weakly in $D_E[0, \infty)$ to the structured coalescent $\{\alpha(t) : t \geq 0\}$ with $Q$-matrix $Q$ given by

$$Q_{\alpha, \beta} = \begin{cases} -\sum_{i \in S} \left\{ \alpha_i \frac{M_i}{2} + \frac{1}{c_i} \left( \alpha_i^2 \right) \right\} & \text{if } \beta = \alpha \\ \frac{\alpha_i}{c_i} \left( \alpha_i^2 \right) & \text{if } \beta = \alpha - \epsilon^i + \epsilon^j (j \neq i) \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

for $\alpha, \beta \in E$.

Note that $4N_{m_{ij}}$ is, up to the constant $c_i$, twice the number of immigrants into subpopulation $i$ from subpopulation $j$ per generation. The assumption that $4N_{m_{ij}} (j \neq i)$ monotonically increases as a function of $N$ is equivalent to assuming that the number of migrants between any two subpopulations per generation increases as the subpopulation sizes increase. This assumption ensures, by the monotone convergence theorem, that $\forall i \in S : M_i = \lim_{N \to \infty} (4N_{m_i})$, so that $c_iM_i$ is twice the limiting number of migrants from (or into) subpopulation $i$ per generation. Assumption (ii) requires that the number of migrants from each subpopulation per generation remains bounded as the subpopulations become large.

Proof of Theorem 2.1: In order to prove the theorem, we show that the finite-dimensional distributions of the process $\alpha_N([2Nt])$ converge to those of the structured coalescent as $N \to \infty$.
and that the family of processes \( \{a_N([2N-])\}_{N \in \mathbb{N}} \) is relatively compact. As \( E \) is countable, and hence separable, Theorem 7.8 in Ethier and Kurtz (1986) then yields the weak convergence of \( a_N([2N-]) \) to the structured coalescent, as \( N \to \infty \).

**Convergence of the finite-dimensional distributions.** From the description of the model it is clear that the ancestral process \( \alpha_N \) is a (multi-dimensional) discrete-time Markov chain. We will calculate its transition matrix \( P_N \) and we will show that, for fixed \( t \geq 0 \), the transition matrix of \( \alpha_N \) over \([2Nt]\) generations, \( P_N^{[2Nt]} \), converges entry-wise to the transition matrix of the structured coalescent over time \( t \), \( e^{tQ} \) (where \( Q \) is the infinitesimal generator of the structured coalescent, given by equation (2.2)), as \( N \to \infty \). This means that the one-dimensional distributions of \( \alpha_N([2N-]) \) converge to those of the structured coalescent, as \( N \to \infty \). Because of the Markov character of both the ancestral process and the structured coalescent and as \( E \) is countable, this is easily seen to be equivalent to the convergence of the finite-dimensional distributions.

Taking a step of one generation backward in time, we first have to take a migration step and subsequently a reproduction step. In order to find the transition probability of the ancestral process from state \( \alpha \in E \) to state \( \beta \in E \) in one generation,

\[
P_N(\beta|\alpha) := P(\alpha_N(\tau + 1) = \beta|\alpha_N(\tau) = \alpha),
\]

we calculate separately the probabilities \( P_N^{(m)}(\beta|\alpha) \) and \( P_N^{(r)}(\beta|\alpha) \) of a transition of the ancestral process from \( \alpha \) to \( \beta \) in, respectively, one backward migration step or one backward reproduction step.

In the migration step, \( m_kN_k \) individuals from subpopulation \( k \) (\( k \in S \)) move, backward in time, to another subpopulation. These migrants are a random sample without replacement from subpopulation \( k \). The number of possible ways to draw \( m_kN_k \) migrants without replacement from among the \( N_k \) individuals in subpopulation \( k \) is \( \binom{N_k}{m_kN_k} \). The individuals counted in \( \alpha \) (the present value of the ancestral process) are the ancestors of our initial sample of \( n_0 \) individuals; we call them individuals "belonging to" \( \alpha \). If an individual belonging to \( \alpha \) is a migrant, the ancestral process may change value. Migrants drawn from outside \( \alpha \) do not affect the ancestral process. The probability \( R_N^{(m)}(\alpha) \) that more than one ancestor in \( \alpha \) is a migrant is the probability that at least two of the migrants are drawn from \( \alpha \) and, counting the number of possible ways to do this, satisfies

\[
R_N^{(m)}(\alpha) \leq \sum_{k \in S} \binom{\alpha_k}{2} \binom{N_k - 2}{m_kN_k - 2} + \sum_{k \in S} \binom{\alpha_k}{1} \binom{N_k - 1}{m_kN_k - 1} \sum_{l \neq k} \binom{\alpha_l}{1} \binom{N_l - 1}{m_lN_l - 1}
\]

\[
\leq \sum_{k \in S} \alpha_k^2m_k^2 + \sum_{k \in S} \alpha_km_k \sum_{l \neq k} \alpha_l m_l
\]

\[
= \left( \sum_{k \in S} \alpha_km_k \right)^2.
\]
Chapter 2: THE STRUCTURED COALESCENT

The probability that exactly one individual in \( \alpha \) migrates backward in time from subpopulation \( i \) to subpopulation \( j(\neq i) \) while all other migrants are drawn from outside \( \alpha \), is

\[
\frac{\binom{\alpha}{1} \cdot \binom{N_i - \alpha_i}{m_{ij}N_i - 1} \cdot \binom{N_i - m_{ij}N_i - \alpha_i + 1}{m_iN_i - m_{ij}N_i} \cdot \prod_{k \neq i} \binom{N_k - \alpha_k}{m_kN_k}}{\binom{N_i}{m_{ij}N_i} \cdot \binom{N_i - m_{ij}N_i - \alpha_i + 1}{m_iN_i - m_{ij}N_i}} = \alpha_i m_{ij} \cdot \frac{N_i}{N_i - m_iN_i - \alpha_i + 1} \prod_{k \in S} \prod_{a=0}^{\alpha_k - 1} \frac{N_k - m_kN_k - a}{N_k - a},
\]

where \( \prod_{a=0}^{-1} \equiv 1 \). In that event, the backward migration step changes the value of the ancestral process to \( \alpha - \epsilon^i + \epsilon^j \). The ancestral process also takes this value when there are several migrations of individuals belonging to \( \alpha \) which, except for one migration from subpopulation \( i \) to subpopulation \( j \), all compensate each other. Denoting by \( R_N^{(m)}(\alpha, \beta) \) the probability that the backward migration step changes the value of the ancestral process from \( \alpha \) to \( \beta \) and more than one individual in \( \alpha \) are migrants, the transition probability in one backward migration step is given by

\[
P_N^{(m)}(\beta|\alpha) = \begin{cases} 
1 - \sum_{i \in S} \alpha_i m_i \frac{N_i}{N_i - m_iN_i - \alpha_i + 1} \prod_{k \in S} \prod_{a=0}^{\alpha_k - 1} \frac{N_k - m_kN_k - a}{N_k - a} - \sum_{\gamma \neq \alpha} R_N^{(m)}(\alpha, \gamma) & \text{if } \beta = \alpha \\
\alpha_i m_{ij} \frac{N_i}{N_i - m_iN_i - \alpha_i + 1} \prod_{k \in S} \prod_{a=0}^{\alpha_k - 1} \frac{N_k - m_kN_k - a}{N_k - a} + R_N^{(m)}(\alpha, \alpha - \epsilon^i + \epsilon^j) & \text{if } \beta = \alpha - \epsilon^i + \epsilon^j \quad (j \neq i) \\
R_N^{(m)}(\alpha, \beta) & \text{otherwise,}
\end{cases}
\]

where

\[
\sum_{\beta \neq \alpha} R_N^{(m)}(\alpha, \beta) \leq R_N^{(m)}(\alpha) \leq \left( \sum_{k \in S} \alpha_k m_k \right)^2. \tag{2.5}
\]

In the backward reproduction step, all individuals in subpopulation \( i \) choose their parent at random, independently and uniformly from among the \( N_i \) individuals which made up subpopulation \( i \) just before reproduction \( (i \in S) \). Two individuals in subpopulation \( i \) choose the same parent with probability \( 1/N_i \). If \( \alpha \) is the present value of the ancestral process and exactly two of the \( \alpha_i \) ancestors counted in subpopulation \( i \) share their parent, while all other ancestors in \( \alpha \) have distinct parents, the value of the ancestral process changes to \( \alpha - \epsilon^i \). The probability of this transition is

\[
P_N^{(r)}(\alpha - \epsilon^i|\alpha) = \binom{\alpha_i}{2} \frac{1}{N_i} - R_N^{(r)}(\alpha, \alpha - \epsilon^i), \tag{2.6}
\]
where $R_N^{(r)}(\alpha, \alpha - e^i)$ is a non-negative term arising from the possibility that two or more pairs of individuals belonging to $\alpha$ each share a parent. As

$$\sum_{i \in S} P_N^{(r)}(\alpha - e^i | \alpha) = P_N^{(r)} \{ \text{exactly one pair of individuals in } \alpha \text{ share a parent} \}$$

and

$$\sum_{i \in S} \left( \begin{array}{c} \alpha_i \\ 2 \end{array} \right) \frac{1}{N_i} = \sum_{\nu = 1}^{\infty} \nu P_N^{(r)} \{ \text{exactly } \nu \text{ pairs of individuals in } \alpha \text{ each share a parent} \},$$

it follows that

$$\sum_{i \in S} R_N^{(r)}(\alpha, \alpha - e^i) = \sum_{\nu = 2}^{\infty} \nu P_N^{(r)} \{ \text{exactly } \nu \text{ pairs of individuals in } \alpha \text{ each share a parent} \}. \quad (2.7)$$

We denote by $V = \sum_{i \in S} \left( \begin{array}{c} \alpha_i \\ 2 \end{array} \right)$ the number of pairs of individuals in $\alpha$ in which both individuals belong to the same subpopulation and hence can have a common parent. As there are at most $n_0$ individuals in $\alpha$ ($n_0$ is the size of the initial sample), we know that $V \leq \left( \begin{array}{c} n_0 \\ 2 \end{array} \right)$. Hence, since $\forall i \in S : N_i = 2c_i N \geq 2N$, the probability $R_N^{(r)}(\alpha)$ that two or more pairs of individuals belonging to $\alpha$ each share a parent is bounded by

$$R_N^{(r)}(\alpha) \leq \left( \begin{array}{c} n_0 \\ 2 \end{array} \right) \left( \frac{1}{2N} \right)^2 \leq \frac{n_0^4}{32N^2}, \quad (2.8)$$

and from equation (2.7) it follows that the quantities $R_N^{(r)}(\alpha, \alpha - e^i)$ in equation (2.6) are bounded in terms of the probability $R_N^{(r)}(\alpha)$ by

$$2R_N^{(r)}(\alpha) \leq \sum_{i \in S} R_N^{(r)}(\alpha, \alpha - e^i) \leq \left( \begin{array}{c} n_0 \\ 2 \end{array} \right) R_N^{(r)}(\alpha). \quad (2.9)$$

Denoting by $R_N^{(r)}(\alpha, \beta)$ the probability that the backward reproduction step changes the value of the ancestral process from $\alpha$ to $\beta \notin \{ \alpha \cup \{ \alpha - e^i : i \in S \} \}$, we have that

$$R_N^{(r)}(\alpha) = \sum_{\beta \notin \{ \alpha \cup \{ \alpha - e^i : i \in S \} \}} R_N^{(r)}(\alpha, \beta). \quad (2.10)$$

Combining the above, the transition probability of the ancestral process in one backward reproduction step is given by

$$P_N^{(r)}(\beta | \alpha) = \begin{cases} 
1 - \sum_{i \in S} \left( \begin{array}{c} \alpha_i \\ 2 \end{array} \right) + \sum_{i \in S} R_N^{(r)}(\alpha, \alpha - e^i) - \sum_{\gamma \notin \{ \alpha \cup \{ \alpha - e^i : i \in S \} \}} R_N^{(r)}(\alpha, \gamma) & \text{if } \beta = \alpha \\
\frac{1}{N_i} \left( \begin{array}{c} \alpha_i \\ 2 \end{array} \right) - R_N^{(r)}(\alpha, \alpha - e^i) & \text{if } \beta = \alpha - e^i \\
R_N^{(r)}(\alpha, \beta) & \text{otherwise,} 
\end{cases} \quad (2.11)$$
where
\[
\sum_{\beta \neq \alpha} R_N^{(\gamma)}(\alpha, \beta) \leq \left\{ \left( \begin{array}{c} n_0 \\ 2 \end{array} \right) + 1 \right\} R_N^{(\gamma)}(\alpha).
\]

As migration and reproduction operate independently, the one-generation transition probabilities of the ancestral process, (2.3), are found from the transition probabilities in one backward migration step and one backward reproduction step as
\[
P_N(\beta|\alpha) = \sum_{\gamma} P_N^{(m)}(\gamma|\alpha) P_N^{(r)}(\beta|\gamma).
\]
In matrix notation:
\[
P_N = P_N^{(m)} \cdot P_N^{(r)},
\]
where \(P_N\), \(P_N^{(m)}\) and \(P_N^{(r)}\) are the transition matrices in one generation, one backward migration step and one backward reproduction step, respectively (for example, the entries of \(P_N\) are \((P_N)_{\alpha,\beta} = P_N(\beta|\alpha)\)). Denoting by \(I\) the identity matrix, with entries
\[
I_{\alpha,\beta} = \delta_{\alpha,\beta} = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{otherwise}, \end{cases}
\]
equation (2.4) for \(P_N^{(m)}(\beta|\alpha)\) can be written as
\[
P_N^{(m)} = I + \frac{1}{2N} Q_N^{(m)} + R_N^{(m)},
\]
where \(Q_N^{(m)}\) and \(R_N^{(m)}\) respectively denote the matrices with entries
\[
\left( Q_N^{(m)} \right)_{\alpha,\beta} = \begin{cases} - \sum_{i \in S} \frac{N_i}{N_i - m_i N_i - \alpha_i + 1} \prod_{k \in S} \prod_{a=0}^{a_k-1} \frac{N_k - m_k N_k - a}{N_k - a} & \text{if } \beta = \alpha \\ \frac{\alpha_i(2Nm_i)}{N_i - m_i N_i - \alpha_i + 1} \prod_{k \in S} \prod_{a=0}^{a_k-1} \frac{N_k - m_k N_k - a}{N_k - a} & \text{if } \beta = \alpha - \epsilon^i + \epsilon^j (j \neq i) \\ 0 & \text{otherwise} \end{cases}
\]
and
\[
\left( R_N^{(m)} \right)_{\alpha,\beta} = \begin{cases} - \sum_{\gamma \neq \alpha} R_N^{(m)}(\alpha, \gamma) & \text{if } \beta = \alpha \\ R_N^{(m)}(\alpha, \beta) & \text{otherwise}. \end{cases}
\]
From equation (2.11) it is seen that the matrix \(P_N^{(r)}\) can be partitioned similarly as
\[
P_N^{(r)} = I + \frac{1}{2N} Q_N^{(r)} + R_N^{(r)},
\]
where \(Q^{(r)}\) is the matrix with entries
\[
\left( Q^{(r)} \right)_{\alpha,\beta} = \begin{cases} - \sum_{i \in S} \frac{1}{c_i} \left( \begin{array}{c} \alpha_i \\ 2 \end{array} \right) & \text{if } \beta = \alpha \\ \frac{1}{c_i} \left( \begin{array}{c} \alpha_i \\ 2 \end{array} \right) & \text{if } \beta = \alpha - \epsilon^i \\ 0 & \text{otherwise} \end{cases}
\]
(independent of $N$) and $R_N^{(r)}$ is the matrix with entries
\[ (R_N^{(r)})_{\alpha, \beta} = \begin{cases} \sum_{\gamma \in S} R_N^{(r)}(\alpha, \alpha - \varepsilon^i) - \sum_{\gamma \in S} R_N^{(r)}(\alpha, \gamma) & \text{if } \beta = \alpha \\ -R_N^{(r)}(\alpha, \alpha - \varepsilon^i) & \text{if } \beta = \alpha - \varepsilon^i \\ R_N^{(r)}(\alpha, \beta) & \text{otherwise.} \end{cases} \]

Note that when the number of subpopulations is infinite, the state space of the ancestral process, $E$, and hence the above matrices, are infinite. We consider the following norm on these matrices:
\[
\|A\| := \sup_{\alpha \in E} \sum_{\beta \in E} |A_{\alpha, \beta}|
\]
where the matrix $A$ has entries $A_{\alpha, \beta}$ ($\alpha, \beta \in E$). Using assumptions (i) and (ii), restricting to $N > \frac{1}{4} \sup_{i \in S} M_i + \frac{1}{4} n_0$ and using the bounds in (2.5), (2.8), (2.9) and (2.10), all the matrices above have finite norm:
\[
\begin{align*}
\|Q_N^{(m)}\| &\leq \frac{1}{4} n_0 N \left( 1 - \frac{1}{4N} \sup_{i \in S} M_i - \frac{n_0}{2N} \right) < \infty \\
\|Q^{(r)}\| &\leq 2 \left( \frac{n_0}{2} \right) \\
\|R_N^{(m)}\| &\leq 2n_0^2 \\
\|Q_N^{(r)}\| &\leq \frac{n_0^3}{32 n_0^2}.
\end{align*}
\]
while all three transition matrices have norm 1. Hence the product of any two of the above matrices exists (with finite entries) and has finite norm (bounded above by the product of the norms of these two matrices). Substituting equations (2.13) and (2.14) into equation (2.12), the one-generation transition matrix of the ancestral process $\alpha_N$ can be written as
\[
P_N = I + \frac{1}{2N} (Q_N + \Delta_N),
\]
where
\[
\begin{align*}
Q_N &= Q_N^{(m)} + Q^{(r)} \\
\Delta_N &= 2N \left( R_N^{(m)} + R_N^{(r)} + R_N^{(m)} R_N^{(r)} + \frac{1}{2N} Q_N^{(m)} R_N^{(r)} + \frac{1}{2N} R_N^{(m)} Q^{(r)} + \frac{1}{4N^2} Q_N^{(m)} Q^{(r)} \right).
\end{align*}
\]
As only a finite number of the $\alpha_i$ ($i \in S$) are non-zero and with $Q$ denoting the infinitesimal generator of the structured coalescent, given by equation (2.2), where $M_{ij}$ and $M_i$ are defined by the assumptions (i) and (ii), it is clear that
\[
\lim_{N \to \infty} Q_N = Q, \text{ entry-wise,}
\]
i.e.
\[
\forall \alpha, \beta \in E : \lim_{N \to \infty} (Q_N)_{\alpha, \beta} = Q_{\alpha, \beta}.
\]
We prove that for fixed $t \geq 0$, the transition matrix of the ancestral process $\alpha_N$ over $[2Nt]$ generations converges entry-wise to the transition matrix of the structured coalescent over time $t$:

$$\lim_{N \to \infty} P_N^{[2Nt]} = e^t Q,$$ entry-wise. \hfill (2.19)

First note that

$$\|Q\| \leq n_0 \sup_{i \in S} M_i + 2 \left( \begin{array}{c} n_0 \\ 2 \end{array} \right) < \infty,$$

so that $e^t Q := \sum_{v=0}^{\infty} \frac{t^v Q^v}{v!}$ exists, as

$$\forall \alpha, \beta \in E : \left| \left( e^t Q \right)_{\alpha \beta} \right| \leq \sum_{v=0}^{\infty} \frac{t^v \|Q^v\|_{\alpha \beta}}{v!} \leq \sum_{v=0}^{\infty} \frac{t^v \|Q\|^v}{v!} = e^t \|Q\| < \infty.$$

Using equation (2.16), the matrix $P_N^{[2Nt]}$ can be written as

$$P_N^{[2Nt]} = \left\{ I + \frac{1}{2N} (Q_N + \Delta_N) \right\}^{2Nt}$$

$$= \sum_{v=0}^{[2Nt]} \frac{[2Nt]}{v} \left( \frac{1}{2N} \right)^v (Q_N + \Delta_N)^v$$

$$= \sum_{v=0}^{2Nt} \frac{[2Nt]!}{(2N)^v} \frac{[2Nt]!}{[2Nt] - v + 1} \frac{[2Nt]!}{[2Nt] - v + 1} \frac{(Q_N + \Delta_N)^v}{v!}.$$

Hence for $\alpha, \beta \in E$:

$$\left( P_N^{[2Nt]} \right)_{\alpha \beta} = \sum_{v=0}^{\infty} a_{v, N}$$ \hfill (2.20)

with

$$a_{v, N} = I_{v \leq [2Nt]} \cdot \frac{[2Nt]!}{(2N)^v} \frac{[2Nt]!}{[2Nt] - v + 1} \frac{[2Nt]!}{[2Nt] - v + 1} \frac{(Q_N + \Delta_N)^v}{v!}$$ \hfill (2.21)

where

$$I_{v \leq [2Nt]} = \begin{cases} 1 & \text{if } v \leq [2Nt] \\ 0 & \text{otherwise.} \end{cases}$$

As $N \to \infty$ and for fixed $v$, the first factor in the right-hand side of (2.21) converges to 1 and the second factor converges to $t^v$. We focus on the third factor. Assumptions (i) and (ii) imply that

$$\sup_{i \in S} m_i = O(1/N), \text{ as } N \to \infty.$$ \hfill (2.22)

Hence it follows from (2.5) that

$$\|P_N^{(m)}\| = 2n_0^2 \left( \sup_{i \in S} m_i \right)^2 = O\left( \frac{1}{N^2} \right),$$

while (2.15) yields that also

$$\|P_N^{(r)}\| = O\left( \frac{1}{N^2} \right).$$

Restricting to

$$N \geq \frac{1}{2} \sup_{i \in S} M_i + n_0,$$ \hfill (2.23)
we have that 
\[ \|Q_N^{(m)}\| \leq n_0 \sup_{i \in S} M_i \cdot \frac{1}{1 - \frac{1}{N} \sup_{i \in S} M_i - \frac{n_0}{2N}} \leq 2n_0 \sup_{i \in S} M_i, \]
which is independent of \( N \). Substituting this information into equation (2.17), it follows that
\[ \|\Delta_N\| = O\left(\frac{1}{N}\right), \text{ as } N \to \infty. \] (2.24)

For \( N \) subject to (2.23), we have in addition that
\[ \|Q_N\| \leq \|Q_N^{(m)}\| + \|Q^{(r)}\| \leq C, \] (2.25)

where
\[ C := \left(2n_0 \sup_{i \in S} M_i + \|Q^{(r)}\|\right) \] (2.26)
is finite and independent of \( N \). Expanding the matrix \((Q_N + \Delta_N)^\nu\), it takes the form
\[ (Q_N + \Delta_N)^\nu = Q_N^\nu + A_N \] (2.27)
with
\[ \|A_N\| = O\left(\frac{1}{N}\right), \]
\( A_N \) being a finite sum of products of matrices \( Q_N \) and \( \Delta_N \), where each product contains at least one factor \( \Delta_N \). A fortiori, \( A_N = O\left(\frac{1}{N}\right) \), entry-wise, i.e.
\[ \forall \alpha, \beta \in E : (A_N)_{\alpha, \beta} = O\left(\frac{1}{N}\right). \] (2.28)

Denoting by \( V \) the matrix with entries
\[ V_{\alpha, \beta} = \begin{cases} \sum_{i \in S} \left\{ \alpha_i M_i + \frac{1}{\alpha_i} \begin{pmatrix} \alpha' \\ 2 \end{pmatrix} \right\} & \text{if } \beta = \alpha \\ \alpha_i M_{ij} & \text{if } \beta = \alpha - e^j + e^i \ (j \neq i) \\ \frac{1}{\alpha_i} \begin{pmatrix} \alpha' \\ 2 \end{pmatrix} & \text{if } \beta = \alpha - e^i \\ 0 & \text{otherwise,} \end{cases} \]
we have for \( N \) subject to (2.23) that
\[ \forall \alpha, \beta \in E : |(Q_N)^\nu_{\alpha, \beta}| \leq V_{\alpha, \beta}. \]

As \( V \) does not depend on \( N \) and \( \|V\| \leq 2n_0 \sup_{i \in S} M_i + n_0(n_0 - 1) < \infty \), it follows from (2.18) by the dominated convergence theorem that
\[ \lim_{N \to -\infty} (Q_N)^\nu_{\alpha, \beta} = \lim_{N \to -\infty} \sum_{\gamma_1, \ldots, \gamma_{\nu-1}} (Q_N)_{\alpha, \gamma_1} \cdot (Q_N)^{\gamma_1, \gamma_2, \ldots, \gamma_{\nu-1}, \gamma_{\nu-1}, \beta} \\
= \sum_{\gamma_1, \ldots, \gamma_{\nu-1}} Q_{\alpha, \gamma_1} \cdot Q_{\gamma_1, \gamma_2} \cdot \ldots \cdot Q_{\gamma_{\nu-1}, \beta} \\
= (Q^\nu)_{\alpha, \beta}. \]
From (2.27) and (2.28) we obtain that
\[
\lim_{N \to \infty} (Q_N + \Delta_N)^v = Q^v, \text{ entry-wise.}
\]
Hence the quantities \( a_{v,N} \) defined by equation (2.21), have as their respective limits
\[
\lim_{N \to \infty} a_{v,N} = \frac{t^v(Q^v)_{\alpha,\beta}}{v!},
\]
(2.29)
for all \( v \in \mathbb{N} \). Using (2.24) and (2.25), we have for \( N \) sufficiently large that \( \|Q_N\| \leq C \) and \( \|\Delta_N\| \leq 1 \), and hence that
\[
|a_{v,N}| \leq \frac{t^v\|Q_N + \Delta_N\|^v}{v!} \leq \frac{t^v(C+1)^v}{v!}, \forall v \in \mathbb{N}.
\]
As \( \sum_{v=0}^{\infty} \frac{t^v(C+1)^v}{v!} = e^{t(C+1)} < \infty \), it follows from (2.20) and (2.29) by the dominated convergence theorem that
\[
\forall \alpha, \beta \in E : \lim_{N \to \infty} \left( P_{[2N]}^{[2N]} \right)_{\alpha,\beta} = \sum_{v=0}^{\infty} \frac{t^v(Q^v)_{\alpha,\beta}}{v!} = e^{tQ} \left( \alpha,\beta \right).
\]
Because of the Markov character of both the ancestral process \( \alpha_N \) and the structured coalescent and because \( E \) is countable, this result implies that the finite-dimensional distributions of the process \( \{\alpha_N([2N]) ; t \geq 0\} \) converge to those of the structured coalescent \( \{\alpha(t) ; t \geq 0\} \) as \( N \to \infty \).

Relative compactness. We regard \( E \) as a subspace of \( L^\infty \) (where \( L \) is the set of the real numbers), endowed with the norm
\[
\|x\| = \sup_{t \in \mathbb{S}} |x_t|,
\]
(2.31)
for \( x = (x_t)_{t \in \mathbb{S}} \in L^\infty \). With this norm, \( E \) is complete, as every Cauchy-sequence in \( E \) is in the long run constant. According to Corollary 7.4 in Ethier and Kurtz (1986), the relative compactness of \( \{\alpha_N([2N])\} \) is guaranteed if we prove the following two conditions:\(^1\)
(a) For every \( \eta > 0 \) and \( t \geq 0 \), there exists a compact set \( \Gamma_{\eta,t} \subset E \) such that
\[
\liminf_{N \to \infty} P\{\alpha_N([2N]) \in \Gamma_{\eta,t}\} \geq 1 - \eta.
\]
(b) For every \( \eta > 0 \) and \( T > 0 \), there exists \( \delta > 0 \) such that
\[
\limsup_{N \to \infty} P\{w'(\alpha_N([2N]), \delta, T) \geq \eta\} \leq \eta,
\]
(2.32)
where \( w' \) is the modulus of continuity:
\[
w'(\alpha_N([2N]), \delta, T) = \inf_{\{t_i\}} \max_i \sup_{s.t \in \{t_i \}} \|\alpha_N([2N]) - \alpha_N([2N])\|,
\]
where \( \{t_i\} \) ranges over all partitions of the form \( 0 = t_0 < t_1 < \cdots < t_{k-1} < T \leq t_k \) with \( \min_{1 \leq i \leq k} (t_i - t_{i-1}) > \delta \) and \( k \geq 1 \).

\(^1\)The condition (a) stated here is slightly stronger than that of Ethier and Kurtz. In particular, we have used that \( \Gamma_{\eta,t} \) is a subset of the set \( \Gamma_{\eta,t}^\infty \) in condition (a) of Ethier and Kurtz.
Chapter 2: THE STRUCTURED COALESCENT

To verify condition (a), fix \( t^0 \geq 0 \) and \( 0 < \eta < 1 \). With the norm (2.31), \( E \) is not compact if the number of subpopulations is infinite. However, as \( E \) is countable, there exists a finite (and hence compact) set \( \Gamma_{\eta,t} \subset E \) such that

\[
P\{\alpha(t) \in \Gamma_{\eta,t}\} \geq 1 - \eta.
\]  

(2.33)

Since we have proved that the one-dimensional distributions of the ancestral process \( \alpha_N([2N\cdot]) \) converge to those of the structured coalescent \( \alpha(\cdot) \) as \( N \to \infty \),

\[
P\{\alpha(t) \in \Gamma_{\eta,t}\} = \lim_{N \to \infty} P\{\alpha_N([2Nt]) \in \Gamma_{\eta,t}\}.
\]

Combining this with (2.33), condition (a) is proved.

Condition (b) essentially requires that the ancestral process \( \alpha_N \), which is a pure jump process, does not jump too quickly, in the limit as \( N \to \infty \), so that a partition \( \{t_i\} = \{t^N_i\} \) can be found which satisfies the requirements set out in the definition of \( w' \) and which contains all the jump times of the ancestral process. To prove this, we couple the jumps of \( \alpha_N \) to those of a process with a higher jump rate but whose inter-jump times are identically distributed and which is therefore easier to handle.

For \( N \in \mathbb{N}_0 \) we denote

\[
p_N := \frac{1}{2N}(C + 1),
\]

where \( C \) is given by equation (2.26). In the remainder of the proof we restrict to \( N \) sufficiently large so that \( p_N < 1 \) and \( \|Q_N + \Delta_N\| \leq (C + 1) \) (see equations (2.24) and (2.25)). For each such \( N \), we define the discrete-time Markov chain \((Z_N, \xi_N) \equiv \{(Z_N(\tau), \xi_N(\tau)) ; \tau = 0, 1, 2, \ldots \}\) to have state space \( \mathbb{N} \times E \) and transition probabilities

\[
P\{(Z_N(\tau + 1), \xi_N(\tau + 1)) = (j, \beta)|(Z_N(\tau), \xi_N(\tau)) = (i, \alpha)\} = \begin{cases} 1 - p_N & \text{if } j = i \text{ and } \beta = \alpha \\ p_N - \sum_{\gamma \in E: \gamma \neq \alpha} P_N(\gamma|\alpha) & \text{if } j = i + 1 \text{ and } \beta = \alpha \\ P_N(\beta|\alpha) & \text{if } j = i + 1 \text{ and } \beta \neq \alpha \\ 0 & \text{otherwise,} \end{cases}
\]

(2.34)

where \( P_N(\beta|\alpha) \) is the transition probability of the ancestral process \( \alpha_N \) from \( \alpha \) to \( \beta \) in one generation. Using equation (2.16) and the restrictions made on \( N \), we have \( \forall \alpha \in E \) that

\[
\sum_{\gamma \in E: \gamma \neq \alpha} P_N(\gamma|\alpha) = \frac{1}{2N} \sum_{\gamma \neq \alpha} (Q_N + \Delta_N)_{\alpha,\gamma} \leq p_N,
\]

so that \((Z_N, \xi_N)\) is well defined, as the elements of its transition matrix, (2.34), are non-negative and, when summing over \((j, \beta) \in \mathbb{N} \times E\), add to one. From (2.34), it is clear that the marginal
distribution of \( \xi_N \) is that of the ancestral process \( \alpha_N \), so that

\[
P\{w'(\alpha_N([2N]), \delta, T) \geq \eta \} = P\{w'(\xi_N([2N]), \delta, T) \geq \eta \}
\]  
(2.35)

for every \( \eta > 0 \) and \( T > 0 \). The process \( Z_N \) jumps with probability \( p_N \), every generation; at each jump, its value increases by one. The construction is such that every time \( \xi_N \) jumps, \( Z_N \) jumps as well.

We denote by \( 0 = \rho_N^0 < \rho_N^1 < \ldots \) the jump times of the process \( (Z_N, \xi_N) \) and by \( \tau_i^N := \rho_i^N - \rho_{i-1}^N \) \((i \in \mathbb{N}_0)\) its inter-jump times. Because of the Markov character of \( (Z_N, \xi_N) \), the \( \tau_i^N \) are mutually independent. As the probability of a jump of \( (Z_N, \xi_N) \) is \( p_N \) every generation, each \( \tau_i^N \) is geometrically distributed with mean \( 1/p_N \).

Now fix \( \eta > 0 \) and \( T > 0 \). If for some \( J \in \mathbb{N}_0 \) and for some \( \delta > 0 \),

\[
\rho_i^N \geq 2NT \quad \text{and} \quad \tau_i^N > 2N\delta \quad \text{for } i = 1, \ldots, J,
\]
then denoting \( k_N := \min \{i : \rho_i^N \geq 2NT\} \) we have that \( 1 \leq k_N \leq J \) and the partition

\[
t_i^N := \frac{\rho_i^N}{2N} \quad \text{for } i = 0, \ldots, k_N
\]

satisfies

\[
0 = t_0^N < t_1^N < \cdots < t_{k_N}^N < T \leq t_{k_N}
\]

and

\[
t_i^N - t_{i-1}^N > \delta \quad \text{for } i = 1, \ldots, k_N;
\]
as the process \( (Z_N, \xi_N) \) is constant between the jump times \( \rho_i^N \), we have in that case that \( \xi_N([2N]) \) is constant on each interval \([t_{i-1}^N, t_i^N)\) \((i = 1, \ldots, k_N)\), so that

\[
w'(\xi_N([2N]), \delta, T) = 0.
\]

Hence for every \( J \in \mathbb{N}_0 \) and \( \delta > 0 \):

\[
P\{w'(\xi_N([2N]), \delta, T) < \eta \} \geq P\{\rho_i^N \geq 2NT \quad \text{and} \quad \tau_i^N > 2N\delta \quad \text{for } i = 1, \ldots, J\}. \tag{2.36}
\]

Thus in order to prove condition (b) it is sufficient to find \( J \in \mathbb{N}_0 \) and \( \delta > 0 \) such that

\[
\liminf_{N \to \infty} P\{\rho_i^N \geq 2NT \quad \text{and} \quad \tau_i^N > 2N\delta \quad \text{for } i = 1, \ldots, J\} \geq 1 - \eta.
\]

Now

\[
P\{\rho_i^N \geq 2NT \quad \text{and} \quad \tau_i^N > 2N\delta \quad \text{for } i = 1, \ldots, J\}
\]

\[
= P\{\rho_i^N \geq 2NT|\tau_i^N > 2N\delta \quad \text{for } i = 1, \ldots, J\} P\{\tau_i^N > 2N\delta \quad \text{for } i = 1, \ldots, J\}
\]

\[
= P\{\rho_i^N \geq 2NT|\tau_i^N > 2N\delta \quad \text{for } i = 1, \ldots, J\} \left(P\{\tau_i^N > 2N\delta\}\right)^J, \tag{2.37}
\]
because the $\tau^N_i$ ($i = 1, \ldots, J$) are i.i.d. As $\rho^N_j = \sum_{i=1}^{J} \tau^N_i$, it is seen either by direct calculation or by a straightforward correlation argument that
\[
P\{\rho^N_j \geq 2NT | \tau^N_i > 2N\delta \text{ for } i = 1, \ldots, J\} \geq P\{\rho^N_j \geq 2NT\}.
\] (2.38)

Since $\rho^N_j$ is the time of the $J$th jump of the process $(Z_N, \xi_N)$ and $Z_N$ counts the number of jumps of $(Z_N, \xi_N)$,
\[
P\{\rho^N_j > 2NT\} = P\{Z_N([2NT]) - Z_N(0) < J\}.
\] (2.39)

Combining (2.35), (2.36), (2.37), (2.38) and (2.39) we have that
\[
P\{w'(\alpha_N([2N:]), \delta, T) < \eta\} \geq P\{Z_N([2NT]) - Z_N(0) < J\} \left( P\left\{\frac{\tau^N}{2N} > \delta \right\} \right)^J,
\] (2.40)

for every $J \in \mathbb{N}_0$ and $\delta > 0$. Because the distribution of $\tau^N_i$ is geometric with mean $1/p_N$ and as \(\lim_{N \to \infty} (2Np_N) = C + 1\), $\tau^N/(2N)$ converges in distribution, as $N \to \infty$, to an exponentially distributed random variable, $X$, with mean $1/(C + 1)$. As the probability of a jump of the process $(Z_N, \xi_N)$ is $p_N$ every generation and $Z_N([2NT]) - Z_N(0)$ is the number of jumps up to generation $[2NT]$, the distribution of $Z_N([2NT]) - Z_N(0)$ is binomial with parameters $[2NT]$ and $p_N$, so that as $N \to \infty$, $Z_N([2NT]) - Z_N(0)$ converges in distribution to a random variable, $Z$, which is Poisson distributed with mean equal to $\lim_{N \to \infty} ([2NT]p_N) = T(C + 1)$. Using (2.40), we obtain that
\[
\lim_{N \to \infty} \inf P\{w'(\alpha_N([2N:]), \delta, T) < \eta\} \geq P\{Z < J\} \left( P\{X > \delta \} \right)^J,
\] (2.41)

which can be made arbitrarily close to 1 by first choosing $J$ sufficiently large and subsequently choosing $\delta$ small. Thus (2.32) follows and the proof is completed. \(\Box\)

### 2.3 Coalescence times

The structured coalescent, being a time-homogeneous continuous-time Markov chain, lends itself very well to analytical study. In population genetics applications, one is often interested in the time since the most recent common ancestor of a sample from the population, or, in some cases, in the time since a sample of $k$ individuals descended from $k - 1$ distinct ancestors, as these times are closely related to genetic variation and differentiation. These "coalescence times" can, in principle, be calculated from the structured coalescent by a first-step argument. In this section we derive systems of linear equations for the moment-generating function, the mean and the second moment of the time since the most recent common ancestor of a sample of individuals (Subsection 1), and of the time since $k$ individuals descended from $k - 1$ ancestors (Subsection 2).

In Chapter 3 these equations will be solved for samples of size two under a variety of models of population structure. The resulting coalescence times will be used in Chapter 4 to study genetic
differentiation among subpopulations under specific population structures. Chapter 5 contains some findings on coalescence times and on subpopulation differentiation which are valid under fairly general conditions on the population structure.

2.3.1 The coalescence time of a sample from the population

Throughout this work, the time since the most recent common ancestor of a sample of individuals will be called their coalescence time. Every sample from the population corresponds to an element of \( \mathbb{N}^S \), by counting the number of individuals sampled from each subpopulation. The coalescence time of a sample with configuration \( \alpha \in \mathbb{N}^S \) will be denoted by \( T_\alpha \). It is the time until the structured coalescent \( \alpha(\cdot) \), starting from \( \alpha(0) = \alpha \), enters into the absorbing set of states \( \{e^i; i \in S\} \). Denoting \( |\gamma| = \sum_{i \in S} \gamma_i \), for every \( \gamma = (\gamma_i)_{i \in S} \in \mathbb{N}^S \),

\[
T_\alpha = \inf\{t \geq 0; |\alpha(t)| = 1\},
\]

where the structured coalescent \( \alpha(\cdot) \) is started from \( \alpha(0) = \alpha \).

The structured coalescent is the same limiting process as that studied by Notohara (1990), and is closely related to an interacting particle system analyzed by Shiga (1980). Denoting by \( X_1 \) and \( X_2 \) two independent Markov chains in continuous time, both generated by the matrix of the scaled migration rates, \( M = (M_{ij})_{i,j \in S} \), where\( \forall i \in S: M_{ii} := -M_{ii} \), and assuming that these Markov chains are irreducible, Shiga (1980) introduced the following classification:

Case I

\[
P\left\{ \int_0^\infty I_{\{X_1(t) = X_2(t)\}} dt = +\infty \mid (X_1(0), X_2(0)) = (i_1, i_2) \right\} = 1 \quad \text{for all} \quad (i_1, i_2) \in S \times S
\]

Case II

\[
P\left\{ \int_0^\infty I_{\{X_1(t) = X_2(t)\}} dt < +\infty \mid (X_1(0), X_2(0)) = (i_1, i_2) \right\} = 1 \quad \text{for all} \quad (i_1, i_2) \in S \times S
\]

Case III

\[
0 < P\left\{ \int_0^\infty I_{\{X_1(t) = X_2(t)\}} dt = +\infty \mid (X_1(0), X_2(0)) = (i_1, i_2) \right\} < 1 \quad \text{for all} \quad (i_1, i_2) \in S \times S,
\]

where \( \forall (x_1, x_2) \in S \times S: \)

\[
I_{\{x_1 = x_2\}} = \begin{cases} 
1 & \text{if } x_1 = x_2 \\
0 & \text{otherwise.}
\end{cases}
\]

Under the assumption that \( X_1 \) and \( X_2 \) are irreducible, the above three cases exhaust all possibilities (Shiga 1980). The theorem below (Notohara 1990) gives a necessary and sufficient condition under which any finite sample from the population has with probability one a common ancestor within a finite time. Notohara's result is analogous to Lemma 3.2 in Shiga (1980); both results are the same if all subpopulations have the same size.

**Theorem 2.2 (Notohara 1990)** Assume the Markov chain generated by the matrix of the scaled migration rates, \( M = (M_{ij})_{i,j \in S} \), is irreducible. The following statements are equivalent:
(i) \( M \) satisfies the condition of Case I.

(ii) \( \forall \alpha \in \mathbb{N}^S \text{ with } |\alpha| < \infty : P\{T_{\alpha} < \infty\} = 1 \).

**Proof:** see Notohara (1990) and the proof of Lemma 3.2 in Shiga (1980).

We will use the term migration chain to refer to the Markov chain generated by the scaled migration rates (the continuous-time Markov chain \( X = \{X(t) : t \geq 0\} \) whose Q-matrix has off-diagonal elements \( M_{ij} (i, j \in S) \)). Notohara (1990) mentioned that the condition of Case I is always true if \( S \) is finite and the migration chain is irreducible. Thus for the case of a finite number of subpopulations, the following theorem is obtained, which we also prove directly:

**Theorem 2.3** Assume the population is divided into a finite number of subpopulations. If the migration chain is irreducible, then \( \forall \alpha \in \mathbb{N}^S : \)

\[
P\{T_\alpha < \infty\} = 1 \tag{2.42}
\]

and

\[
ET_\alpha < \infty. \tag{2.43}
\]

**Proof:** Assume \( S \) is finite and the migration chain is irreducible. We first prove (2.42) and (2.43) for samples of size two. Because the structured coalescent with initial sample size \( n_0 = 2 \) is not irreducible, we construct a modified process \( \{a^*(t) : t \geq 0\} \) with Q-matrix \( Q^* = \left(Q_{\alpha, \beta}^*\right) \) given by

\[
Q_{\alpha, \beta}^* = Q_{\alpha, \beta} \quad \text{if } |\alpha| = 2,
\]

\[
Q_{e_i, \beta}^* = \begin{cases} 
-\frac{1}{c}, & \text{if } \beta = e^i \\
\frac{1}{c}, & \text{if } \beta = 2e^i \\
0, & \text{otherwise}
\end{cases}
\]

\((i \in S)\). Starting from \( \alpha \in \mathbb{N}^S \text{ with } |\alpha| = 2 \), the "modified structured coalescent" \( \{a^*(t) : t \geq 0\} \) has the same distribution as the structured coalescent \( \{\alpha(t) : t \geq 0\} \) until coalescence occurs. Because of the assumptions, \( a^*(\cdot) \) is irreducible and has finite state space, and is hence positive recurrent (for example, Grimmett and Stirzaker 1992). Denoting by \( T_{\alpha, e^i} \) the first passage time of \( a^*(\cdot) \) through the state \( e^i \), starting from \( a^*(0) = \alpha \ (|\alpha| = 2) \), and fixing \( i_0 \in S \), it follows that \( \forall \alpha \in \mathbb{N}^S \text{ with } |\alpha| = 2 : \)

\[
P\{T_{\alpha, e^{i_0}} < \infty\} = 1
\]

and

\[
ET_{\alpha, e^{i_0}} < \infty.
\]
Chapter 2: THE STRUCTURED COALESCENT

As the coalescence time $T_\alpha$ of a sample with configuration $\alpha$ satisfies

$$T_\alpha \overset{d}{=} \min_{i \in S} T_{\alpha_i, e} \leq T_{\alpha, e^u},$$

statements (i) and (ii) hold, for every $\alpha \in N^S$ with $|\alpha| = 2$.

Now consider a sample with configuration $\alpha \in N^S$ with $|\alpha| > 2$. Labelling the individuals in the sample as $1, \ldots, |\alpha|$ and denoting by $T(u, v)$ the coalescence time of individuals $u$ and $v$, we have that

$$T_\alpha = \max\{T(u, v); u, v = 1, \ldots, |\alpha| \text{ with } u \neq v\} \leq \sum_{u=1}^{[\alpha]} \sum_{v=u+1}^{[\alpha]} T(u, v).$$

As individuals $u$ and $v$ form a sample of size two, the above argument gives

$$P\{T(u, v) < \infty\} = 1$$

and

$$ET(u, v) < \infty,$$

for every $u, v \in \{1, \ldots, |\alpha|\}$ with $u \neq v$. Hence

$$P\{T_\alpha = \infty\} \leq \sum_{u=1}^{[\alpha]} \sum_{v=u+1}^{[\alpha]} P\{T(u, v) = \infty\} = 0$$

and

$$ET_\alpha \leq \sum_{u=1}^{[\alpha]} \sum_{v=u+1}^{[\alpha]} ET(u, v) < \infty.$$  

\square

Theorem 2.3 states that, in the case of a finite number of subpopulations, the irreducibility of the migration chain is sufficient to guarantee that the coalescence time of every sample from the population is finite with probability one and has finite expectation. In Theorem 5.5 (Chapter 5) we will prove that under Assumption 2.1, the irreducibility of the migration chain is also a necessary condition for (i) or (ii) to hold, if the number of subpopulations is finite. The same theorem will in addition give an equivalent condition in terms of graph theory.

The distribution of the coalescence time of a sample from the population can, in principle, be obtained from the structured coalescent. The distribution of a non-negative random variable $Y$ is uniquely determined by its Laplace transform, which is the function $\mathbb{E}[e^{-sY}]$ of $s \geq 0$. This Laplace transform, evaluated in $s$, is also the moment-generating function of $Y$, evaluated in $-s$. We will abuse terminology and refer to the Laplace transform, $\mathbb{E}[e^{-sY}]$, as the moment-generating function. From the structured coalescent we will by a first-step argument derive systems of linear equations for the "moment-generating function" of the coalescence time of a sample from the population. In Chapter 3 we will solve this system of equations to find the distribution of the coalescence time of a pair of individuals under various specific population structures.
Theorem 2.4 The following system of linear equations holds for the moment-generating function of the coalescence time of a finite sample from the population:

\[
\left( \sum_{i \in S} \left\{ \frac{1}{c_i} \left( \frac{\alpha_i}{2} \right) + \alpha_i \frac{M_i}{2} \right\} + s \right) E[e^{-sT_\alpha}] - \sum_{i \in S} \sum_{j \neq i} \alpha_i \frac{M_{ij}}{2} E[e^{-s(T_\alpha-e_i+e_j)}] \\
- \sum_{i \in S} \frac{1}{c_i} \left( \frac{\alpha_i}{2} \right) E[e^{-sT_\alpha-e_i}] = 0 \text{ if } 2 \leq |\alpha| < \infty,
\]

(2.44)

\[E[e^{-sT_\alpha}] = 1 \text{ if } |\alpha| = 1.\]

Proof: The statement for \(|\alpha| = 1\) is obvious since in that case \(T_\alpha = 0\).

For \(\alpha \in N^S\) with \(2 \leq |\alpha| < \infty\), denote by \(X_\alpha\) the waiting time of the structured coalescent in the state \(\alpha\). The distribution of \(X_\alpha\) is exponential with mean \(|Q_{\alpha,\alpha}|^{-1}\) (where \(Q\) is the infinitesimal generator of the structured coalescent). Because of the Markov character of the structured coalescent, conditioning on the first step taken by the embedded jump chain of the structured coalescent yields that

\[E[e^{-sT_\alpha}] = E[e^{-sX_\alpha}] \sum_{\beta \neq \alpha} \frac{Q_{\alpha,\beta}}{|Q_{\alpha,\alpha}|} E[e^{-sT_\beta}]\]

\[= \frac{|Q_{\alpha,\alpha}|}{|Q_{\alpha,\alpha}| + s} \sum_{\beta \neq \alpha} \frac{Q_{\alpha,\beta}}{|Q_{\alpha,\alpha}|} E[e^{-sT_\beta}]\]

\[= \frac{1}{|Q_{\alpha,\alpha}| + s} \sum_{\beta \neq \alpha} Q_{\alpha,\beta} E[e^{-sT_\beta}],\]

so that

\[(|Q_{\alpha,\alpha}| + s) E[e^{-sT_\alpha}] - \sum_{\beta \neq \alpha} Q_{\alpha,\beta} E[e^{-sT_\beta}] = 0.\]

Substituting (2.2) into this equation, equation (2.44) is obtained. \(\square\)

In the following two theorems, similar equations are obtained for the mean and the second moment of the coalescence time of a sample from the population.

Theorem 2.5 (Notohara 1990) If \(ET_\alpha < \infty\) for every \(\alpha \in N^S\) with \(|\alpha| < \infty\), the following system of linear equations holds for the mean coalescence time of a finite sample from the population:

\[\sum_{i \in S} \left\{ \frac{1}{c_i} \left( \frac{\alpha_i}{2} \right) + \alpha_i \frac{M_i}{2} \right\} ET_\alpha - \sum_{i \in S} \sum_{j \neq i} \alpha_i \frac{M_{ij}}{2} ET_\alpha - e_i + e_j\]

\[- \sum_{i \in S} \frac{1}{c_i} \left( \frac{\alpha_i}{2} \right) ET_\alpha - e_i = 1 \text{ if } 2 \leq |\alpha| < \infty,\]

\[ET_\alpha = 0 \text{ if } |\alpha| = 1.\]
Chapter 2: THE STRUCTURED COALESCENT

Proof: The result follows from a first-step argument (Notohara 1990) similar to the one in the proof of Theorem 2.4, or by differentiating the equations of Theorem 2.4 with respect to $s$ and subsequently taking the limit as $s \to 0$. □

Theorem 2.6 If $E[T^2_\alpha] < \infty$ for every $\alpha \in \mathbb{N}^S$ with $|\alpha| < \infty$, the following system of linear equations holds for the second moment of the coalescence time of a finite sample from the population:

$$
\sum_{i \in S} \left( \frac{1}{c_i} \left( \begin{array}{c} s \\ 2 \\ \end{array} \right) + \alpha_i \frac{M_i}{2} \right) E[T^2_\alpha] - \sum_{i \in S, j \neq i} \alpha_i \frac{M_{ij}}{2} E[T^2_{\alpha - e_i + e_j}] \\
- \sum_{i \in S} \frac{1}{c_i} \left( \begin{array}{c} s \\ 2 \\ \end{array} \right) E[T^2_{\alpha - e_i}] = 2ET_\alpha \text{ if } 2 \leq |\alpha| < \infty,
$$

$$
E[T^2_\alpha] = 0 \text{ if } |\alpha| = 1.
$$

Proof: The result follows from a first-step argument similar to the one in the proof of Theorem 2.4 or by differentiating the equations of Theorem 2.4 twice with respect to $s$ and then taking the limit as $s \to 0$. □

2.3.2 The "k to k - 1" coalescence time of a sample from the population

We call the "$k$ to $k - 1$" coalescence time of a sample of size $k \geq 2$ the time since the $k$ sampled individuals descended from $k - 1$ distinct ancestors. If a sample from the population has configuration $\alpha \in \mathbb{N}^S$ with $|\alpha| = k$, its "$k$ to $k - 1$" coalescence time, denoted by $T^\alpha_k$, is the time until the structured coalescent $\{\alpha(t) : t \geq 0\}$, starting from $\alpha(0) = \alpha$, enters into the set of states $\{\gamma \in \mathbb{N}^S : |\gamma| = k - 1\}$. Formally, for every $\alpha \in \mathbb{N}^S$ with $2 \leq |\alpha| < \infty$:

$$
T^\alpha_k = \inf\{t \geq 0 : |\alpha(t)| = |\alpha| - 1\},
$$

where the structured coalescent $\alpha(\cdot)$ is started from $\alpha(0) = \alpha$.

For the "$k$ to $k - 1$" coalescence time of a sample of size $k$, systems of linear equations can be derived, similar to the ones found in the previous subsection for the coalescence time of a sample.

Theorem 2.7 The following system of linear equations holds for the moment-generating function of the "$k$ to $k - 1$" coalescence time of sample of $k$ individuals from the population ($k \geq 2$):

$$
\left( \sum_{i \in S} \left( \frac{1}{c_i} \left( \begin{array}{c} s \\ 2 \\ \end{array} \right) + \alpha_i \frac{M_i}{2} \right) + s \right) E[e^{-sT^\alpha_k}] - \sum_{i \in S, j \neq i} \alpha_i \frac{M_{ij}}{2} E[e^{-sT^\alpha_k - e_i + e_j}] = \sum_{i \in S} \frac{1}{c_i} \left( \begin{array}{c} s \\ 2 \\ \end{array} \right)
$$

for $\alpha \in \mathbb{N}^S$ with $|\alpha| = k$. 
Proof: The result follows from the same first-step argument as in the proof of Theorem 2.4. □

**Theorem 2.8** If $ET'_{\alpha} < \infty$ for every $\alpha \in N^S$ with $|\alpha| = k$, the following system of linear equations holds for the mean "$k$ to $k-1$" coalescence time of a sample of size $k$ ($k \geq 2$):

$$\sum_{i \in S} \left\{ \frac{1}{c_i} \binom{\alpha_i}{2} + \alpha_i \frac{M_i}{2} \right\} ET'_{\alpha} - \sum_{i \in S} \sum_{j \neq i} \alpha_i \frac{M_{ij}}{2} ET'_{\alpha - e_i + e_j} = 1,$$

for $\alpha \in N^S$ with $|\alpha| = k$.

Proof: The result follows by a first-step argument or by differentiating the equations of Theorem 2.7 with respect to $s$ and taking the limit as $s \downarrow 0$. □

**Theorem 2.9** If $E[T'_{\alpha}^2] < \infty$ for every $\alpha \in N^S$ with $|\alpha| = k$, the following system of linear equations holds for the second moment of the "$k$ to $k-1$" coalescence time of a sample of $k$ individuals ($k \geq 2$):

$$\sum_{i \in S} \left\{ \frac{1}{c_i} \binom{\alpha_i}{2} + \alpha_i \frac{M_i}{2} \right\} E[T'_{\alpha}^2] - \sum_{i \in S} \sum_{j \neq i} \alpha_i \frac{M_{ij}}{2} E[T'_{\alpha - e_i + e_j}] = 2ET'_{\alpha},$$

for $\alpha \in N^S$ with $|\alpha| = k$.

Proof: The result follows by a first-step argument or by differentiating the equations of Theorem 2.7 twice with respect to $s$ and taking the limit as $s \downarrow 0$. □
Chapter 3

GENEALOGY UNDER
VARIOUS MODELS OF
POPULATION STRUCTURE

This chapter is concerned with the coalescence time of a sample of two individuals under various specific population structures. The models considered are the symmetric island model (Section 1), the finite and infinite stepping-stone models in one and two dimensions (Section 2) and some less symmetric structures (Section 3), namely a general two-population model (Subsection 3.1), a "continental island model" (Subsection 3.2) and a four-population model with geographic barrier (Subsection 3.3). For each model, the moment-generating function, mean and second moment of the coalescence time of a pair of individuals are calculated. For the symmetric island model we also derive the density function of the coalescence time of a pair of individuals. To illustrate the effect of population structure on genealogy, the density function of the coalescence time of two individuals from a single subpopulation under the symmetric island model is compared to that of two individuals from a panmictic population of the same size as the total subdivided population. For the symmetric island model, we will also calculate the mean coalescence time and the mean "3 to 2" coalescence time of a sample of three individuals.

The results obtained in this chapter will be used in Chapter 4 to study the genetic differentiation among subpopulations under the various models of population structure considered.

Whereas most analytical results obtained in the remainder of this thesis assume only that the structured coalescent, given by equation (2.2), is an appropriate description of the genealogy of a sample from the population, our interpretation of some of the results assumes that, with a single time-scaling, coalescence rates in different populations are inversely proportional to the sizes of these populations. In particular, whenever in this thesis we compare a subdivided population with
a panmictic population of the same total size, we assume that, with the same time-scaling as used for the subdivided population, any two individuals in the panmictic population coalesce at rate \((\sum_{i \in S} c_i)^{-1}\), where \(c_i\) is defined by equation (2.2) (that is, \(c_i^{-1}\) is the coalescence rate of a pair of individuals in subpopulation \(i\)). When studying populations with unequal-sized subpopulations, we assume that the \(c_i\) (\(i \in S\)) in equation (2.2) correspond to the relative subpopulation sizes (that is, if subpopulation \(i\) has size \(N_i\) for every \(i \in S\), then \(N_i/N_j = c_i/c_j\) for every \(i, j \in S\)). One example of a reproductive model which meets these assumptions is the neutral Wright-Fisher model described in Section 2.1.

### 3.1 The symmetric island model

![Figure 3.1: Symmetric island model with \(n = 4\) subpopulations. Each circle stands for a subpopulation and the arrows indicate migration.](image)

The symmetric island model, first formulated by Wright (1931), is by far the simplest and most commonly studied model of population structure. Under this model the population is divided into \(n\) equal-sized subpopulations \((n \geq 2\) and finite\) with the same migration rate between any two subpopulations.

Under the appropriate assumptions and time-scale (see, for example, Section 2.1 and Theorem 2.1), the genealogy of a sample from this population is well described by the structured coalescent \(\{\alpha(t): t \geq 0\}\) with infinitesimal generator \(Q\) given by equation (2.2), where \(S = \{1, \ldots, n\}\), \(c_i = 1\) for \(i = 1, \ldots, n\), where all \(M_i\) are equal, say \(M_i = M\) for \(i = 1, \ldots, n\), and where \(M_{ij} = M/(n - 1)\) for \(j \neq i\). With this notation, the scaled migration rate of an individual to any specific other subpopulation is \(M/(2(n - 1))\). Under the model set out in Section 2.1, time-scaling is in units of \(2N\) generations, where \(2N\) is the number of haploid individuals in each subpopulation, and \(M = \lim_{N \to \infty} (4Nm)\), where \(m\) is the proportion of each subpopulation that is replaced by immigrants every generation. So under that model, \(M\) is (in the coalescent approximation) twice the number of individuals migrating from (or to) each subpopulation every generation. The migration rate from subpopulation \(i\) to subpopulation \(j\) in discrete time is \(q_{ij} = m/(n - 1)\) for
Chapter 3: GENEALOGY

\[ i, j = 1, \ldots, n \text{ with } j \neq i. \]

Drawing a sample of size two from the population, denoting by \( T_0 \) the coalescence time of two individuals from the same subpopulation and by \( T_1 \) that of two individuals from different subpopulations, the system of equations given in Theorem 2.4 or, equivalent for a sample of two individuals, Theorem 2.7, reduces to

\[
\begin{align*}
(1 + M + s)E[e^{-sT_0}] - ME[e^{-sT_1}] &= 1 \\
(M + (n-1)s)E[e^{-sT_1}] - ME[e^{-sT_0}] &= 0.
\end{align*}
\]

The solutions of this system of linear equations are

\[
\begin{align*}
E[e^{-sT_0}] &= \frac{M + (n-1)s}{M + (nM + n - 1)s + (n-1)s^2} \quad (3.1) \\
E[e^{-sT_1}] &= \frac{M}{M + (nM + n - 1)s + (n-1)s^2} \quad (3.2)
\end{align*}
\]

(Hudson 1990). These or similar results have been found before without consideration of the structured coalescent, by numerous authors, including Maruyama (1970; in which a small correction is needed, as was pointed out by Latter 1973), Nei (1975), Griffiths (1981), Nagylaki (1983) and Crow and Aoki (1984).

The mean and the second moment of the coalescence time of a pair of individuals are found by differentiation of (3.1) and (3.2) or by solving the systems of equations given in Theorems 2.5 and 2.6 (or Theorems 2.8 and 2.9). We obtain:

\[
\begin{align*}
ET_0 &= n \quad (3.3) \\
ET_1 &= n + \frac{n-1}{M} \quad (3.4)
\end{align*}
\]

(Hudson 1990, Notohara 1990, Hey 1991) and

\[
\begin{align*}
E[T_0^2] &= 2n^2 + 2\frac{(n-1)^2}{M} \quad (3.5) \\
E[T_1^2] &= 2n^2 + 2\frac{(n-1)(2n-1)}{M} + 2\frac{(n-1)^2}{M^2} \quad (3.6)
\end{align*}
\]

(equivalent to results in Hey 1991).

For the symmetric island model it is possible to calculate the probability density function of the coalescence time of a pair of individuals explicitly. A partial fraction expansion of the moment-generating function of \( T_0 \), equation (3.1), shows that the distribution of \( T_0 \) is a mixture of two exponential distributions:

\[
E[e^{-sT_0}] = A_1E[e^{-sZ_1}] + A_2E[e^{-sZ_2}], \quad (3.7)
\]

where \( Z_i \) is exponentially distributed with mean \( 1/\lambda_i \), with

\[
\begin{align*}
\lambda_1 &= \frac{nM + n - 1 - \sqrt{D}}{2(n-1)} \\
\lambda_2 &= \frac{nM + n - 1 + \sqrt{D}}{2(n-1)}
\end{align*}
\]
and where

\[ D = (nM + n - 1)^2 - 4(n - 1)M \]

\[ A_1 = \frac{\lambda_2 - 1}{\lambda_2 - \lambda_1} \]

\[ A_2 = \frac{\lambda_1 - 1}{\lambda_1 - \lambda_2} \]

(Griffiths 1981). From (3.7), the density of \( T_0 \) is easily found:

\[
 f_{T_0}(t) = A_1 \lambda_1 e^{-\lambda_1 t} + A_2 \lambda_2 e^{-\lambda_2 t}
 = e^{-\left(1 + \frac{2M}{n-1}\right)\frac{t}{2(n-1)}} \left\{ \cosh \left( \frac{t \sqrt{D}}{2(n-1)} \right) - \frac{(n-2)M + n-1}{\sqrt{D}} \sinh \left( \frac{t \sqrt{D}}{2(n-1)} \right) \right\} . \tag{3.8}
\]

To calculate the density of \( T_1 \), we note that the coalescence time of two individuals in different subpopulations is the time \( T_r \) until these two individuals are present in a single subpopulation for the first time, plus the coalescence time of two individuals in the same subpopulation:

\[ T_1 \overset{d}{=} T_r + T_0. \]

Because of the Markov character of the structured coalescent, both times are independent, so that their probability density functions satisfy

\[ f_{T_1} = f_{T_r} \ast f_{T_0}, \]

where \( \ast \) denotes the convolution product. As \( T_r \) is exponentially distributed with mean \( \left( \frac{M}{n-1} \right)^{-1} \), a straightforward calculation yields

\[
 f_{T_1}(t) = A_1 \lambda_1 \int_0^t f_{T_r}(x)e^{-\lambda_1(t-x)} \, dx + A_2 \lambda_2 \int_0^t f_{T_r}(x)e^{-\lambda_2(t-x)} \, dx
 = \frac{2M}{\sqrt{D}} e^{-\left(1 + \frac{2M}{n-1}\right)\frac{t}{2(n-1)}} \sinh \left( \frac{t \sqrt{D}}{2(n-1)} \right) . \tag{3.9}
\]

Results (3.8) and (3.9) were obtained earlier by Takahata (1988) and Nath and Griffiths (1993) in the case of \( n = 2 \) (two subpopulations), but appear to be new in their general form (\( n \) colonies).

To illustrate the effect of population structure on genealogy, we compare in figure 3.2 the density of the coalescence time \( T_0 \) of two individuals from a single subpopulation in a symmetric island model with \( n = 4 \) subpopulations with the density of the coalescence time \( T_{\text{pan}} \) of two individuals in a panmictic population of the same total size \( 2nN \), if the subpopulations in the symmetric island model each contain \( 2N \) haploid individuals. From equation (2.2) with one subpopulation or from Kingman’s coalescent for a panmictic population (see equation (1.1)), the distribution of \( T_{\text{pan}} \) is, in the same time-scaling as used for the island model, exponential with mean \( n \) (see also the assumptions immediately preceding this section). Figure 3.2 shows that for a small migration rate, two individuals from the same subpopulation in the subdivided population (solid line) are more likely to have a very short coalescence time than two individuals in the panmictic population.
Figure 3.2: The coalescence time of two individuals in a single subpopulation of a subdivided population, compared to that of two individuals in a panmictic population of the same total size.

The solid line is the density function $f_{T_0}(t)$, given by equation (3.8), of the coalescence time $T_0$ of two individuals in a single subpopulation under a symmetric island model with $n = 4$ subpopulations and a scaled migration rate of $M = 0.5$. The dashed line is the density function $f_{T_{\text{pan}}}(t)$ of the coalescence time $T_{\text{pan}}$ of two individuals in a panmictic population of the same total size (and with the same time-scaling). The distribution of $T_{\text{pan}}$ is exponential with mean $n$. The distribution of $T_0$ is a mixture of two exponential distributions.

(dashed line). This is because the subpopulation is much smaller than the total population, so that the probability that two individuals share their parent in the previous generation is larger in a single subpopulation of a structured population than it is in a large panmictic population. Once one of the individuals' lineages moves to a different subpopulation, however, coalescence cannot occur until the two lineages are present in a single subpopulation again, which is likely to take a long time if the migration rate is small. This explains the longer and thicker tail of the density of $T_0$, compared to that of $T_{\text{pan}}$. In fact, $f_{T_0}(t)/f_{T_{\text{pan}}}(t) \to \infty$ as $t \to \infty$. These observations also explain why the variance of $T_0$ (obtained from results (3.3) and (3.5); see also Hey 1991) is much larger than that of $T_{\text{pan}}$ and increases as the migration rate decreases:

$$\text{Var}(T_0) = n^2 + 2 \frac{(n-1)^2}{M} > n^2 = \text{Var}(T_{\text{pan}}).$$

However, it is interesting to note that both distributions have the same mean:

$$ET_0 = ET_{\text{pan}} = n.$$
In the next sections we will see that this is in fact the case for various, though not all, population structures. In Chapter 5, two sets of graph-theoretic conditions on the population structure will be given under which the mean coalescence time of two individuals from any single subpopulation is independent of the migration rate and equal to that of two individuals from a panmictic population of the same total size. The mean coalescence time of two individuals from different subpopulations in the symmetric island model, given by equation (3.4), is larger than that of two individuals from a panmictic population of the same total size. The variance of $T_1$, 

$$\text{Var}(T_1) = n^2 + 2\frac{(n-1)^2}{M} + \frac{(n-1)^2}{M^2}$$

(obtained from results (3.4) and (3.6); see also Hey 1991), is also larger than that of $T_{\text{pan}}$.

If we let the scaled migration rate $M$ increase to infinity in the moment-generating functions (3.1) and (3.2), we find that

$$\lim_{M \to \infty} E[e^{-sT_0}] = \lim_{M \to \infty} E[e^{-sT_1}] = \frac{1/n}{1/n + s} = E[e^{-sT_{\text{pan}}}],$$

for every $s \geq 0$. By Theorems XIII.1.2 and XIII.1.1 in Feller (1966) it follows that $T_0$ and $T_1$ converge in distribution to $T_{\text{pan}}$, as $M \to \infty$, as one would expect intuitively: as the migration rate becomes very large, the population becomes effectively panmictic. From results (3.3) to (3.6) we see that also

$$\lim_{M \to \infty} ET_0 = \lim_{M \to \infty} ET_1 = n = ET_{\text{pan}}$$

and

$$\lim_{M \to \infty} E[T_0^2] = \lim_{M \to \infty} E[T_1^2] = 2n^2 = E[T_{\text{pan}}^2].$$

If $M = 0$ (no migration), the population is made up of $n$ isolated subpopulations. As each subpopulation is panmictic, the coalescence time of two individuals from the same subpopulation, denoted by $T_0^{(0)}$, has in that case a unit exponential distribution. Letting the scaled migration rate $M$ decrease to zero in equation (3.1), we find

$$\lim_{M \to 0} E[e^{-sT_0}] = \frac{1}{1 + s} = E[e^{-sT_0^{(0)}}],$$

(3.10)

for every $s \geq 0$, so that $T_0$ converges in distribution to $T_0^{(0)}$, as $M \downarrow 0$. However,

$$\lim_{M \to 0} ET_0 = n \neq ET_0^{(0)},$$

(3.11)

which implies that $T_0$ does not converge to $T_0^{(0)}$ in $L_1$, as $M \downarrow 0$. This interesting behaviour as the migration rate tends to zero, equations (3.10) and (3.11), was noted earlier by Nath and Griffiths (1993). Indexing the coalescence time of two individuals from a single subpopulation by the scaled migration rate, $T_0 = T_0^{(M)}$, it follows by Proposition 2.3 in the Appendix of Ethier and Kurtz (1986) that the collection of random variables $\{T_0^{(M)}; M > 0\}$ is not uniformly integrable. The coalescence time of two individuals from different subpopulations is, in the case of isolation.
(\(M = 0\)), infinite with probability one. Taking the limit of equations (3.2) and (3.4) as \(M \downarrow 0\), we have indeed
\[
\lim_{M \to 0} E[e^{-sT_1}] = 0,
\]
for \(s > 0\), and
\[
\lim_{M \to 0} ET_1 = +\infty
\]
(see also Nath and Griffiths 1993). Note also that both
\[
\lim_{M \to 0} \text{Var}(T_0) = +\infty
\]
and
\[
\lim_{M \to 0} \text{Var}(T_1) = +\infty.
\]

Drawing a sample of three individuals from the population, there are, if there are three or more subpopulations, essentially three configurations possible, represented by the states \(3e^1, 2e^1 + e^2\) and \(e^1 + e^2 + e^3\). Recalling the notation \(T_0\) introduced in Section 2.3.2 for the time until the first coalescence event, tracing the ancestral lineages of a sample with configuration \(\alpha\), Theorem 2.8 gives the following system of equations for the mean \("3 to 2"\) coalescence time of the sample:
\[
\begin{align*}
(3 + \frac{3M}{2}) ET_{3e^1} - \frac{3M}{2} ET_{2e^1 + e^2} & = 1 \\
(1 + M(1 - \frac{1}{2(n-1)})^{\alpha}) ET_{2e^1 + e^2} - \frac{M}{2(n-1)} ET_{3e^1} - M(1 - \frac{1}{n-1}) ET_{e^1 + e^2 + e^3} & = 1 \\
\frac{3M}{n-1} ET_{e^1 + e^2 + e^3} - \frac{3M}{n-1} ET_{2e^1 + e^2} & = 1,
\end{align*}
\]
for \(n \geq 3\), with solutions
\[
\begin{align*}
ET_{3e^1} & = \frac{n}{3} - \frac{2(n-1)^2}{3\{2(n-1) + Mn\}} \tag{3.12} \\
ET_{2e^1 + e^2} & = \frac{n}{3} + \frac{2(n-1)}{3\{2(n-1) + Mn\}} \tag{3.13} \\
ET_{e^1 + e^2 + e^3} & = \frac{n}{3} + \frac{(n-1)(n+2)}{3\{2(n-1) + Mn\}} + \frac{2(n-1)^2}{3M\{2(n-1) + Mn\}}.
\end{align*}
\]
Results (3.12) and (3.13) are easily seen to be valid for the case of \(n = 2\) subpopulations as well and have been found earlier, for \(n = 2\) subpopulations, by Takahata (1988) and Nath and Griffiths (1993). In contrast with the mean coalescence time of a sample of two individuals from the same subpopulation, \(ET_{3e^1}\) does depend on the migration rate. In a panmictic population of the same total size and with the same time-scaling as the subdivided population, every pair of ancestral lineages has coalescence rate \(1/n\) (see the structured coalescent, given by equation (2.2), with one subpopulation, or see Kingman's coalescent, given by equation (1.1), with the appropriate re-scaling of time), so that the mean \("3 to 2"\) coalescence time of a sample of three individuals is exponentially distributed with mean \(\frac{n}{3}\). So, the mean \,"3 to 2"\) coalescence time of three individuals from a single subpopulation of the subdivided population is smaller than that of three individuals from a panmictic population of the same total size. Regardless of the configuration of the sample,
the mean “3 to 2” coalescence time of a sample of three individuals from the subdivided population converges to that of three individuals from a panmictic population of the same total size, as the migration rate tends to infinity.

For the mean time since the most recent common ancestor of all three sampled individuals, Theorem 2.5 gives the following equations, for \( n \geq 3 \):

\[
\begin{align*}
(3 + \frac{3M}{2}) \, ET_3 & - \frac{3M}{2} \, ET_{e_1+e_2} = 1 + 3ET_0 \\
(1 + M(1 - \frac{1}{2(n-1)})) \, ET_{e_1+e_2} & - \frac{3M}{2(n-1)} \, ET_3 + M(1 - \frac{1}{n-1}) \, ET_{e_1+e_2} = 1 + ET_1 \\
\frac{3M}{n-1} \, ET_{e_1+e_2} & - \frac{3M}{n-1} \, ET_{e_1+e_2} = 1,
\end{align*}
\]

where \( ET_0 \) and \( ET_1 \) are given by equations (3.3) and (3.4), respectively. The solutions of these equations are

\[
\begin{align*}
ET_3 & = \frac{4n}{3} + \frac{(n-1)^2}{3(2n-1) + Mn} \\
ET_{e_1+e_2} & = \frac{4n}{3} + \frac{(3n-1)(n-1)}{3(2n-1) + Mn} + \frac{2(n-1)^2}{M(2n-1) + Mn} \\
ET_{e_1+e_2+e_3} & = \frac{4n}{3} + \frac{(4n-1)(n-1)}{3(2n-1) + Mn} + \frac{8(n-1)^2}{3M(2n-1) + Mn},
\end{align*}
\]

where (3.14) and (3.15) are valid for \( n \geq 2 \). For the case of \( n = 2 \) colonies, results (3.14) and (3.15) were also found by Nath and Griffiths (1993). Note that \( ET_3 \) also depends on the migration rate. The mean coalescence time of three individuals in a panmictic population of the same total size (and with the same time-scaling as used for the subdivided population) is their mean “3 to 2” coalescence time plus the mean coalescence time of two individuals in the panmictic population, and is hence equal to \( \frac{4n}{3} + n = \frac{4n}{3} \). So, the mean coalescence time of three individuals from a single subpopulation under the symmetric island model is larger than if there were no population subdivision. Note that \( ET_3 \), \( ET_{e_1+e_2} \) and \( ET_{e_1+e_2+e_3} \) all converge to \( \frac{4n}{3} \) as the migration rate tends to infinity.

### 3.2 The stepping-stone models

In natural populations, migration tends to be more frequent between neighbouring subpopulations than between colonies a further geographic distance apart. A model that incorporates this is the stepping-stone model, introduced by Malécot (1948; English translation: 1969) and Kimura (1953). In this model, the subpopulations are situated at the points of a finite or infinite rectangular lattice in one or more dimensions and migration occurs only between adjacent colonies. When the number of colonies in a certain dimension is finite, it is usually assumed that the corresponding subpopulations at both ends are connected by migration, in order to avoid edge-effects. In this section, we calculate the moment-generating function, the mean and the second moment of the time since the most recent common ancestor of a pair of individuals under the finite and infinite stepping-stone models in one and two dimensions.
3.2.1 The circular stepping-stone model

Figure 3.3: Circular stepping-stone model with \( n = 8 \) subpopulations.

The population is divided into \( n \) equal-sized subpopulations (where \( n \) is finite) which are located on a circle. Every generation, a proportion \( m \) of each subpopulation is exchanged for immigrants, half of which come from each of the two neighbouring colonies. This model could resemble colonies around a lake or a mountain, or along the edge of a forest or the shore of an island.

The genealogy of a sample from this population is, under the appropriate assumptions and time-scale (for example, those given in Chapter 2), well described by the structured coalescent with Q-matrix given by equation (2.2), where \( S = \{1, \ldots, n\} \) and

\[
q_{ij} = \begin{cases}
\frac{M}{2} & \text{if } |i - j| \in \{1, n - 1\} \\
0 & \text{otherwise } (j \neq i),
\end{cases}
\]

where \( M \) is the scaled migration rate. Each individual has scaled rate \( M/4 \) of migrating to each of the two neighbouring subpopulations. Under the model of Section 2.1, time-scaling is in units of \( 2N \) generations (where \( 2N \) is the subpopulation size), \( q_{ij} = m/2 \) if \( |i - j| \in \{1, n - 1\} \) and \( q_{ij} = 0 \) otherwise \( (j \neq i) \), and \( M = \lim_{N \to \infty} (4Nm) \), which is twice the number of immigrants per subpopulation per generation.

The distribution of the coalescence time of a pair of individuals under this model depends on their locations only through their distance, \( d \), defined as the number of subpopulations separating the two individuals (a formal definition of distance in a general subdivided population will be given in Chapter 5); \( d \) ranges from zero to \( \lfloor n/2 \rfloor \), the largest integer not larger than \( n/2 \). Denoting by \( T_d \) the coalescence time of two individuals at distance \( d \), the equations given in Theorem 2.4 thus
reduce to recursive equations in $d$:

\[
(1 + M + s)E[e^{-rT_d}] - ME[e^{-rT_1}] = 1
\]  \hspace{1cm} (3.16)

\[
(M + s)E[e^{-rT_d}] - \frac{M}{2}E[e^{-rT_{d-1}}] - \frac{M}{2}E[e^{-rT_{d+1}}] = 0 \quad \text{for } d = 1, \ldots, \lfloor \frac{n}{2} \rfloor - 1
\]  \hspace{1cm} (3.17)

\[
(M + s)E[e^{-rT_{\frac{n}{2}}}] - ME[e^{-rT_{\frac{n}{2} - 1}}] = 0 \quad \text{if } n \text{ is even}
\]  \hspace{1cm} (3.18)

\[
\left(\frac{M}{2} + s\right)E[e^{-rT_{\frac{n}{2} - 1}}] - \frac{M}{2}E[e^{-rT_{\frac{n}{2} + 1}}] = 0 \quad \text{if } n \text{ is odd.}
\]  \hspace{1cm} (3.19)

Note that the coefficients in this system of recursive equations are constant in $d$. The general solution of equation (3.17) is

\[
A_+ \cdot \left(\frac{\lambda_+(s)}{M}\right)^d + A_- \cdot \left(\frac{\lambda_-(s)}{M}\right)^d,
\]  \hspace{1cm} (3.20)

where $\frac{\lambda_+(s)}{M}$ and $\frac{\lambda_-(s)}{M}$ are the solutions of the characteristic equation

\[
\frac{M}{2} \lambda^2 - (M + s)\lambda + \frac{M}{2} = 0,
\]

that is,

\[
\lambda_+(s) = M + s + \sqrt{(2M + s)s}
\]

\[
\lambda_-(s) = M + s - \sqrt{(2M + s)s}.
\]  \hspace{1cm} (3.21)

In (3.20), $A_+$ and $A_-$ are constants, to be found from the boundary conditions, equations (3.16) and (3.18) or (3.19). For example, if the number of subpopulations is even, $A_+$ and $A_-$ are obtained by substituting the general solution, (3.20), into equations (3.16) and (3.18):

\[
\begin{align*}
(1 + M + s)(A_+ + A_-) - (A_+\lambda_+(s) + A_-\lambda_-(s)) &= 1 \\
(M + s) \left\{ A_+ \left(\frac{\lambda_+(s)}{M}\right)^{\frac{n}{2}} + A_- \left(\frac{\lambda_-(s)}{M}\right)^{\frac{n}{2}} \right\} - M \left\{ A_+ \left(\frac{\lambda_+(s)}{M}\right)^{\frac{n}{2} - 1} + A_- \left(\frac{\lambda_-(s)}{M}\right)^{\frac{n}{2} - 1} \right\} &= 0.
\end{align*}
\]

Solving this system of linear equations (or the corresponding system of equations in the case of an odd number of subpopulations) for $A_+$ and $A_-$, and substituting the solutions into (3.20), the solution of the above recursive equations, (3.16) to (3.19), is found: for $d = 0, \ldots, \lfloor n/2 \rfloor$,

\[
E[e^{-rT_d}] = \frac{M^d \left\{ a_+(s)\lambda_+(s)^{\lfloor n/2 \rfloor - d} + a_-(s)\lambda_-(s)^{\lfloor n/2 \rfloor - d} \right\}}{1 + \sqrt{(2M + s)s}} a_+(s)\lambda_+(s)^{\lfloor n/2 \rfloor} + \left(1 - \sqrt{(2M + s)s}\right) a_-(s)\lambda_-(s)^{\lfloor n/2 \rfloor}
\]  \hspace{1cm} (3.22)

where

\[
a_+(s) = \frac{1}{s} \quad \text{if } n \text{ is even}
\]

\[
a_+(s) = \sqrt{(2M + s)s} + s \quad \text{if } n \text{ is odd.}
\]  \hspace{1cm} (3.23)

Maruyama (1970) performed an eigenvector analysis to solve a similar system of recursive equations for the probability of identity by descent (to be defined in Chapter 4) of a pair of genes under the circular stepping-stone model. In Chapter 4 we will show that in the coalescent approximation, the probability of identity by descent of a pair of genes is the Laplace transform of
Chapter 3: GENEALOGY

their coalescence time, evaluated in the scaled mutation rate. Maruyama’s results (with a minor
correction: \([\frac{n+1}{2}] \) should be \([\frac{n}{2}] \), which are complicated, apply to the finite population case.
Taking their limit as the subpopulation size tends to infinity (as is appropriate in the coalescent
approximation), an alternative expression for the moment-generating function of the coalescence
time of two individuals at distance \(d \) is obtained:

\[
E[e^{-sT_d}] = \frac{1}{1 + \sum_{k=0}^{\lfloor n/2 \rfloor} \frac{s + M (1 - \cos \frac{2\pi kd}{n})}{s + M (1 - \cos \frac{2\pi k}{n})}}
\]  

(3.24)

for \(d = 0, \ldots, \lfloor n/2 \rfloor \). Using Simpson’s rule it is easily verified that, indeed, (3.24) solves the
above system of recursive equations, (3.16) to (3.19). As the solution of this system of recursive
equations is unique, results (3.22) and (3.24) are equal. Also in the context of probability of
identity by descent, Malécot (1975) found a (slightly simpler) approximation of (3.22), valid for
\(s << M \).

For the mean and the second moment of the coalescence time of a pair of individuals \(d \) steps
apart, Theorems 2.5 and 2.6 give recursive equations similar to those above ((3.16) to (3.19)).
Solving these equations or, alternatively, differentiating result (3.22) with respect to \(s \), we find:

\[
E[T_d] = n + \frac{d(n - d)}{M}
\]  

(3.25)

\[
E[T_d^2] = 2n^2 + \frac{6d(n - d) + n^2 - 1}{3M} + \frac{d(n - d) \{d(n - d) + n^2 + 1\}}{3M^2}
\]  

(3.26)

for \(d = 0, \ldots, \lfloor n/2 \rfloor \). Note that for \(d = 0 \), result (3.25) gives \(E[T_0] = n \), as was also the case
under the symmetric island model: the mean coalescence time of two individuals from a single
subpopulation is independent of the migration rate and equal to that of two individuals from a
panmictic population of the same total size. The mean coalescence time of two individuals at
distance \(d \), equation (3.25), was calculated earlier by Slatkin (1991) using standard results on
random walks. In fact, all results in this subsection could also have been obtained using results
on first passage times of one-dimensional random walks. This random walk approach will be
illustrated in detail in the subsections on the two-dimensional stepping-stone models.

### 3.2.2 The infinite linear stepping-stone model

Here the population consists of an infinite line of equal-sized subpopulations. Every generation,
a proportion \(m/2 \) of each colony is exchanged with each of its two neighbouring colonies. This
model might be appropriate for a very long array of colonies along a shore or a river-bed.

Under the appropriate assumptions and time-scale, the genealogy of a sample of individuals
from this population is well described by the structured coalescent: in equation (2.2) for its Q-
matrix, put \( S = Z \) and for every \( i, j \in S \) with \( j \neq i \),

\[
c_i = 1
\]

\[
M_{ij} = \begin{cases} 
M/2 & \text{if } |i - j| = 1 \\
0 & \text{otherwise}
\end{cases}
\]

where \( M \) is the scaled migration rate. Under the discrete-time model described in Section (2.1): \( q_{ij} = m/2 \) if \( |i - j| = 1 \) and \( q_{ij} = 0 \) otherwise \( (j \neq i) \), and, if each subpopulation contains \( 2N \) individuals, time-scaling is in units of \( 2N \) generations and \( M = \lim_{N \to \infty} (4N/nm) \).

Theorem 2.4 gives for the moment-generating function of the coalescence time of two individuals at distance \( d \) the same equations as (3.16) and (3.17) in the previous subsection, where equation (3.17) now holds for every \( d \in \mathbb{N}_0 \). The general solution of (3.17) is (3.20). Equation (3.16) provides one boundary condition. As a second boundary condition, we know that as \( d \to \infty \), \( E[e^{-\lambda T_d}] \) remains bounded by 1. Because \( \frac{\lambda_{d+1}(s)}{M} > 1 \) for \( s > 0 \), this implies that \( A_+ = 0 \). The value of \( A_- \) is found subsequently by substituting \( E[e^{-\lambda T_d}] = A_- \left( \frac{\lambda_{d+1}(s)}{M} \right)^d \) into equation (3.16). The resulting expression for the moment-generating function of \( T_d \) is

\[
E[e^{-\lambda T_d}] = \frac{(M + s - \sqrt{(2M + s)s})^d}{M^d \left( 1 + \sqrt{(2M + s)s} \right)}
\]

(3.27) for \( d = 0, 1, 2, \ldots \). This result could also have been obtained by noting that the distance between the two individuals, until it first becomes zero, performs a one-dimensional symmetric random walk and by using results on first passage times (see for example Feller 1966, Section XIV.6).

Note that (3.27) is the limit of the corresponding result for the circular stepping-stone model, (3.22), as the number of colonies on the circle becomes infinitely large. This convergence, which was to be expected intuitively, can also be shown as a consequence of the fact that the first passage time through zero of a symmetric random walk on \( \{0, 1, \ldots, [n/2]\} \), starting at \( d \), converges in distribution to that of a symmetric random walk on \( \mathbb{N} \), starting at \( d \), as \( n \to \infty \) (the detailed argument will be given for the two-dimensional stepping-stone models). Alternatively, one could have noted immediately that the limit as \( n \) tends to infinity of the moment-generating function of the coalescence time of two individuals \( d \) steps apart in a circular stepping-stone model with \( n \) colonies solves equations (3.16) and (3.17) and satisfies the additional boundary condition,
Chapter 3: GENEALOGY

52

lim_{d \to \infty} E[e^{-sT_d}] \leq 1. Since the solution is unique, this limit is the value of $E[e^{-sT_d}]$ for the infinite linear stepping-stone model.

Approximations of result (3.27), valid for $s << M$, were obtained earlier by Malécot (1948) and by Maruyama (1970), in the form of the probability of identity by descent of a pair of genes. As an approximation valid for a circular stepping-stone model with many colonies, Maruyama also obtained a result whose limit as the subpopulation size tends to infinity provides an alternative expression for the moment-generating function of the coalescence time of two individuals $d$ steps apart in the infinite linear stepping-stone model:

$$E[e^{-sT_d}] = \frac{\int_{-\pi}^{\pi} \cos(xd) \frac{1}{M + s - M \cos x} dx}{2\pi + \int_{-\pi}^{\pi} \frac{1}{M + s - M \cos x} dx}$$

(3.28)

for $d = 0, 1, 2, \ldots$. This expression is the limit of result (3.24) for the circular stepping-stone model as $n \to \infty$. That both values, (3.27) and (3.28), are equal follows from the uniqueness of the solution to the equations (3.16) and (3.17) with the additional boundary condition, $\lim_{d \to \infty} E[e^{-sT_d}] \leq 1$.

Differentiating (3.27) with respect to $s$ and taking the limit as $s \downarrow 0$, we find that the mean, and hence also the second moment, of the coalescence time of any pair of individuals are infinite:

$$ET_d = \infty$$

(3.29)

$$E[T_d^2] = \infty$$

(3.30)

for $d = 0, 1, 2, \ldots$. This is effectively a consequence of the null-recurrence of the symmetric random walk in one dimension.

3.2.3 Two-dimensional stepping-stone model on a torus

In this model a two-dimensional array of equal-sized colonies is situated on a torus. Denoting by $K$ the number of subpopulations in one dimension and by $L$ that in the other ($K, L \geq 2$), the total number of subpopulations is $n = K \times L$. Every generation, a proportion $m$ of each subpopulation is exchanged for immigrants, a quarter of which come from each of the four neighbouring subpopulations. The assumption that corresponding colonies on opposite sides of the two-dimensional array are linked by migration, although standard (Maruyama 1970, Malécot 1975, Crow and Aoki 1984, Slatkin 1991, 1993), may be unrealistic. Without this assumption however the model is intractible. This model may still provide a good approximation for a large two-dimensional lattice of colonies.

Choosing $S = \{0, \ldots, K - 1\} \times \{0, \ldots, L - 1\}$, choosing $c_{(i,j)} = 1$ for $i = 0, \ldots, K - 1$ and $j = 0, \ldots, L - 1$ and denoting by $M$ the scaled migration rate, the genealogy of a sample from this population is, under the appropriate assumptions and time-scale, well described by the structured...
coalescent, whose Q-matrix is given by equation (2.2), where for \((i, j), (k, l) \in S\) with \((i, j) \neq (k, l)\):
\[
M_{(i,j),(k,l)} = \begin{cases} 
M/4 & \text{if } (|i-k| \in \{1, K-1\} \text{ and } j = l) \text{ or } (|j-l| \in \{1, L-1\} \text{ and } i = k) \\
0 & \text{otherwise.}
\end{cases}
\]

Every individual has scaled rate \(M/8\) of migrating to each of the four neighbouring subpopulations. In the notation of the discrete-time model of Section (2.1), \(q(i,j|i,k) = m/4\) if either \(|i-k| \in \{1, K-1\}\) and \(j = 1\), or \(|j-l| \in \{1, L-1\}\) and \(i = k\); \(q(i,j|i,k) = 0\) otherwise, for \((k, l) \neq (i, j)\).

Under that model, time-scaling is in units of \(2N\) generations, \(2N\) being the subpopulation size, and \(M = \lim_{N \to \infty} (4Nm)\).

The distribution of the coalescence time of a pair of individuals from this population is a function of the numbers of colonies, \(d_1\) and \(d_2\), separating the two individuals in the respective dimensions of the torus \((d_1 = 0, \ldots, \lfloor K/2 \rfloor; d_2 = 0, \ldots, \lfloor L/2 \rfloor)\). We denote by \(T_{(d_1, d_2)}\) the coalescence time of a pair of individuals at “distance” \((d_1, d_2)\). The equations of Theorems 2.4, 2.5 and 2.6 are, for a sample of two individuals at distance \((d_1, d_2)\), recursive equations in the two recursion indices \(d_1\) and \(d_2\). Rather than solving these recursive equations directly, we found it easier to obtain results using the theory of random walks.

The coalescence time of two individuals \(d_1\) steps apart in the first dimension of the torus and \(d_2\) steps apart in the other dimension is the time \(T_{(d_1, d_2)}^{(r)}\) until the ancestors of the two individuals are for the first time present in a single colony plus the coalescence time of two individuals in a single colony:
\[
T_{(d_1, d_2)} = T_{(d_1, d_2)}^{(r)} + T_{(0,0)}.
\]

Because of the Markov character of the structured coalescent, the two times on the right-hand side are independent, so that
\[
E[e^{-sT_{(d_1, d_2)}}] = E\left[e^{-sT_{(d_1, d_2)}^{(r)}}\right] E\left[e^{-sT_{(0,0)}}\right].
\]

For \(d_1 = d_2 = 0\): \(E[e^{-sT_{(d_1, d_2)}}] = 1\). Assume \((d_1, d_2) \neq (0, 0)\). Until the two individuals are present in a single colony for the first time, their “distance” (seen as a bivariate process with one component for each dimension) performs a symmetric random walk on the rectangular lattice \(\{0, \ldots, \lfloor K/2 \rfloor\} \times \{0, \ldots, \lfloor L/2 \rfloor\}\). The distribution of \(T_{(d_1, d_2)}^{(r)}\) is that of the first passage time through \((0,0)\) of this random walk, starting at \((d_1, d_2) \neq (0, 0)\). To calculate this distribution, it is easier to label the individuals, one and two, and to measure the distance between the two individuals “anti-clockwise” from individual one, in both dimensions of the torus. By “anti-clockwise” we mean the direction corresponding to a transition from location 0 to location 1 in that particular dimension of the torus. The distance between the two individuals (measured anti-clockwise from individual one) thus performs a symmetric random walk on the “\(K \times L\) torus” \(\{0, \ldots, K-1\} \times \{0, \ldots, L-1\}\), where transitions occur at rate \(M\) and where each transition is with
probability 1/4 to each of the four adjacent states. The distribution of \( T_{(d_1, d_2)}^{(r)} \) is the same as that of the first passage time through \((0, 0)\) of the latter random walk, starting at \((d_1, d_2) \neq (0, 0)\). We denote by \((D_1, D_2) \equiv ((D_1, D_2)_\nu : \nu = 0, 1, 2, \ldots )\) the jump chain of the random walk described by the distance between the two individuals, measured anti-clockwise from individual one, and by \(U_{(d_1, d_2)}^{(r)}\) the first passage time of \((D_1, D_2)\) through \((0, 0)\), starting from \((d_1, d_2) \neq (0, 0)\). The relationship between \(T_{(d_1, d_2)}^{(r)}\) and \(U_{(d_1, d_2)}^{(r)}\) is given by

\[
T_{(d_1, d_2)}^{(r)} = \sum_{i=1}^{U_{(d_1, d_2)}^{(r)}} X_i
\]

where the \(X_i\) are mutually independent, exponentially distributed random variables with mean 1/M, which are also independent of \(U_{(d_1, d_2)}^{(r)}\). Denoting by \(F_{(d_1, d_2)(0,0)}\) the generating function of \(U_{(d_1, d_2)}^{(r)}\), that is,

\[
F_{(d_1, d_2)(0,0)}(z) := E\left[z^{U_{(d_1, d_2)}^{(r)}}\right]
\]

it follows that

\[
E\left[e^{-zT_{(d_1, d_2)}^{(r)}}\right] = E \left[ \prod_{i=1}^{U_{(d_1, d_2)}^{(r)}} e^{-zX_i} \right] = EE \left[ \prod_{i=1}^{U_{(d_1, d_2)}^{(r)}} e^{-zX_i} \right]^{U_{(d_1, d_2)}^{(r)}} = E \left[ \prod_{i=1}^{U_{(d_1, d_2)}^{(r)}} E[e^{-zX_i}] \right] = E \left( \frac{M}{M + z} \right)^{U_{(d_1, d_2)}^{(r)}} = F_{(d_1, d_2)(0,0)} \left( \frac{M}{M + z} \right). \tag{3.32}
\]

In order to find \(F_{(d_1, d_2)(0,0)}\), we calculate the probability

\[
P_{(d_1, d_2)(0,0)}^{(\nu)} := P\{(D_1, D_2)_\nu = (0, 0) \mid (D_1, D_2)_0 = (d_1, d_2)\} \tag{3.33}
\]

that the discrete-time random walk \((D_1, D_2)\), starting at \((d_1, d_2)\), is at the origin immediately after the \(\nu\)th transition. Note that \(K\) steps in the same direction (clockwise or anti-clockwise) of the first dimension of the torus, or \(L\) steps in a single direction in the second dimension of the torus, do not alter the position of \((D_1, D_2)\). Assume that among the first \(\nu\) steps of \((D_1, D_2)\), \(z\) steps were in the first dimension of the torus, while \(\nu - z\) steps were in the second dimension. Starting at \(d_1\), \(D_1\) takes the value 0 at step \(x\) if among these \(z\) steps, there were \(d_1 + kK\) more steps clockwise than anti-clockwise, for some \(k \in Z\) (where \(-K\) steps clockwise are to be interpreted as \(+K\) steps
anti-clockwise). Thus

\[ P^{(\nu)}_{(d_1, d_2)(0,0)} = \frac{1}{4^\nu K L} \sum_{\nu=0}^{\infty} \left( \sum_{x=0}^{K-1} \left( 2 \cos \frac{2\nu \pi}{K} \right)^x \cos \frac{2d_1 \nu \pi}{K} \cos \frac{2d_2 \nu \pi}{L} \sum_{w=0}^{L-1} \left( \frac{2 \cos \frac{2\nu \pi}{L}}{K} + \cos \frac{2\nu \pi}{L} \right)^{\nu-x} \cos \frac{2d_2 \nu \pi}{L} \right) \]  

where for \( a \in \mathbb{N} \): \( (\nu)^a \equiv 0 \) if \( b \notin \{0, 1, \ldots, a\} \). The following lemma follows from a slight extension of Theorem 4.3 in Teugels (1986):

**Lemma 3.1** For all \( a, b \in \mathbb{N} \) and \( C \in \mathbb{N}_0 \)

\[ \sum_{x=0}^{\infty} \left( \sum_{\nu=0}^{a} \left( \frac{2 \cos \frac{2\nu \pi}{C}}{C} \right)^x \cos \frac{2b \nu \pi}{C} \right) = \frac{1}{C} \sum_{\nu=0}^{C-1} \left( \frac{2 \cos \frac{2\nu \pi}{C}}{C} \right)^x \cos \frac{2b \nu \pi}{C}. \]

Using this lemma, equation (3.34) gives

\[ P^{(\nu)}_{(d_1, d_2)(0,0)} = \frac{1}{4^\nu K L} \sum_{\nu=0}^{K-1} \cos \frac{2d_1 \nu \pi}{K} \cos \frac{2d_2 \nu \pi}{L} \sum_{x=0}^{L-1} \left( \frac{2 \cos \frac{2\nu \pi}{L}}{K} + \cos \frac{2\nu \pi}{L} \right)^{\nu-x} \cos \frac{2d_2 \nu \pi}{L} \]

\[ = \frac{1}{4^\nu K L} \sum_{\nu=0}^{K-1} \cos \frac{2d_1 \nu \pi}{K} \cos \frac{2d_2 \nu \pi}{L} \sum_{x=0}^{L-1} \left( \frac{2 \cos \frac{2\nu \pi}{L}}{K} + \cos \frac{2\nu \pi}{L} \right)^{\nu-x} \cos \frac{2d_2 \nu \pi}{L}. \]

Introducing the generating function

\[ P_{(d_1, d_2)(0,0)}(z) := \sum_{\nu=0}^{\infty} P^{(\nu)}_{(d_1, d_2)(0,0)} z^\nu, \]

we have for \( 0 \leq z < 1 \):

\[ P_{(d_1, d_2)(0,0)}(z) = \frac{K-1}{K L} \sum_{\nu=0}^{K-1} \cos \frac{2d_1 \nu \pi}{K} \cos \frac{2d_2 \nu \pi}{L} \sum_{x=0}^{L-1} \left( \frac{2 \cos \frac{2\nu \pi}{L}}{K} + \cos \frac{2\nu \pi}{L} \right)^{\nu-x} \cos \frac{2d_2 \nu \pi}{L} \]

\[ = \frac{K-1}{K L} \sum_{\nu=0}^{K-1} \cos \frac{2d_1 \nu \pi}{K} \cos \frac{2d_2 \nu \pi}{L} \sum_{x=0}^{L-1} \left( \frac{2 \cos \frac{2\nu \pi}{L}}{K} + \cos \frac{2\nu \pi}{L} \right)^{\nu-x} \cos \frac{2d_2 \nu \pi}{L}. \]  

(3.35)

The generating function of \( U_{(d_1, d_2)}^{(r)} \) is, for \( (d_1, d_2) \neq (0, 0) \), given by

\[ F_{(d_1, d_2)(0,0)}(z) = \frac{P_{(d_1, d_2)(0,0)}(z)}{P_{(0,0)(0,0)}(z)}. \]

(see for example equation (5.3) in Chapter XV of Feller 1968). Combining this with (3.32) and (3.35), we obtain:

\[ E[e^{-\rho T_{(d_1, d_2)}^{(r)}}] = \frac{P_{(d_1, d_2)(0,0)}(z M + M \rho)}{P_{(0,0)(0,0)}(z M + M \rho)} \]

\[ = \frac{\sum_{\nu=0}^{K-1} \sum_{x=0}^{L-1} \cos \frac{2d_1 \nu \pi}{K} \cos \frac{2d_2 \nu \pi}{L}}{\sum_{\nu=0}^{K-1} \sum_{x=0}^{L-1} \cos \frac{2d_1 \nu \pi}{K} + \cos \frac{2d_2 \nu \pi}{L}}. \]  

(3.36)
which is valid for $d_1 = 0, \ldots, [K/2]$, $d_2 = 0, \ldots, [L/2]$ and for $s > 0$.

For the moment-generating function of the coalescence time of two individuals from a single subpopulation, Theorem 2.4 gives the following equation:

$$(1 + M + s)E[e^{-sT(0,0)}] - \frac{M}{2} E[e^{-sT(1,0)}] - \frac{M}{2} E[e^{-sT(0,1)}] = 1.$$  

Substituting (3.31) into this equation, it follows that

$$E[e^{-sT(0,0)}] = \frac{1}{1 + M + s - \frac{M}{2} E[e^{-sT(1,0)}] - \frac{M}{2} E[e^{-sT(0,1)}]}.$$  

(3.38)

Substituting this and (3.37) into equation (3.31), we find for $d_1 = 0, \ldots, [K/2]$, $d_2 = 0, \ldots, [L/2]$ and $s > 0$:

$$E[e^{-sT(d_1, d_2)}] = \frac{E\left[e^{-sT(1,0)}\right]}{1 + M + s - \frac{M}{2} E[e^{-sT(1,0)}] - \frac{M}{2} E[e^{-sT(0,1)}]}.$$  

(3.39)

$$= \frac{\sum_{v=0}^{K-1} \sum_{w=0}^{L-1} \frac{\cos(2v\pi d_1/K) \cos(2w\pi d_2/L)}{M + s - M \left[\cos(2v\pi/K) + \cos(2w\pi/L)\right]/2} + 1}{KL + \sum_{v=0}^{K-1} \sum_{w=0}^{L-1} \frac{\cos(2v\pi/K) + \cos(2w\pi/L)}{M + s - M \left[\cos(2v\pi/K) + \cos(2w\pi/L)\right]/2}.}.$$  

(3.40)

For a stepping-stone model on a square torus ($K = L$) with possibly different migration rates in the two dimensions of the torus, Maruyama (1970) earlier calculated the probability of identity by descent of a pair of genes by solving a system of recursive equations, without using coalescent techniques or the theory of random walks. Apart from a minor error in Maruyama's results ([3.37] there should be $[n^2/2]$)), Maruyama's value, in the limit of infinite subpopulation size, agrees with our result.

The easiest way to calculate the mean and the second moment of the coalescence time of two individuals at distance $(d_1, d_2)$ is by differentiation of the moment-generating function, equation (3.40). Denoting $S_0 := S \setminus \{(0,0)\}$, we find for $d_1 = 0, \ldots, [K/2]$ and $d_2 = 0, \ldots, [L/2]$:

$$ET(d_1, d_2) = KL + \sum_{(v,w) \in S_0} \frac{1 - \cos(2v\pi d_1/K) \cos(2w\pi d_2/L)}{M \left[1 - \cos(2v\pi/K) + \cos(2w\pi/L)\right]/2}.$$  

(3.41)

and

$$E\left[T^2(d_1, d_2)\right] = 2KL + 2KL \sum_{(v,w) \in S_0} \frac{1 - \cos(2v\pi d_1/K) \cos(2w\pi d_2/L)}{M \left[1 - \cos(2v\pi/K) + \cos(2w\pi/L)\right]/2} + \sum_{(v,w) \in S_0} \frac{1 - \cos(2v\pi d_1/K) \cos(2w\pi d_2/L)}{M \left[1 - \cos(2v\pi/K) + \cos(2w\pi/L)\right]/2} + \sum_{(v,w) \in S_0} \frac{1 - \cos(2v\pi d_1/K) \cos(2w\pi d_2/L)}{M \left[1 - \cos(2v\pi/K) + \cos(2w\pi/L)\right]/2}.$$  

(3.42)

For the mean coalescence time, Slatkin (1991, 1993) obtained (up to minor typographical errors) the same value as (3.41) by solving a matrix equation (using the technique of Maruyama 1970).

Note that equation (3.41) gives for $d_1 = d_2 = 0$: $ET(0,0) = KL$, which is the mean coalescence time of two individuals from a panmictic population of the same total size.
3.2.4 The infinite two-dimensional stepping-stone model

The population consists of a two-dimensional rectangular lattice of equal-sized colonies. The number of colonies in each dimension is infinite. Every generation a proportion $m/4$ of each subpopulation is exchanged with each of the four neighbouring colonies.

Under the appropriate assumptions and time-scale, the genealogy of a sample from this population is described by the structured coalescent with its Q-matrix given by equation (2.2) with $S = \mathbb{Z} \times \mathbb{Z}$, $c_{(i,j)} = 1$ for every $(i,j) \in S$ and, denoting by $M$ the scaled migration rate,

$$M((i,j),(k,l)) = \begin{cases} M/4 & \text{if } |i-k| + |j-l| = 1 \\ 0 & \text{otherwise} \end{cases}$$

for $(i,j),(k,l) \in S$ with $(i,j) \neq (k,l)$. If the particular model of Section 2.1 is assumed, time-scaling is in units of $2N$ generations, where $2N$ is the subpopulation size, and $M = \lim_{N \to \infty} (4Nm)$, which is twice the number of immigrants per subpopulation per generation. In the notation of that model: $q_{(i,j),(k,l)} = m/4$ if $|i-k| + |j-l| = 1$ and $q_{(i,j),(k,l)} = 0$ otherwise, for $(i,j),(k,l) \in S$ with $(i,j) \neq (k,l)$.

Denoting by $T(d_1,d_2)$ the coalescence time of a pair of individuals $d_1$ colonies apart in the first dimension of the lattice and $d_2$ colonies apart in the second dimension ($d_1,d_2 \in \mathbb{N}$), Theorem 2.4
gives the following system of recursive equations for the moment-generating function of $T(d_1,d_2)$:

\[
\begin{align*}
(1 + M + s)E[e^{-sT(o,o)}] - \frac{M}{2} E[e^{-sT(1,0)}] - \frac{M}{2} E[e^{-sT(0,1)}] &= 1 \\
(M + s)E[e^{-sT(1,1)}] - \frac{M}{4} E[e^{-sT(1,1)}] - \frac{M}{4} E[e^{-sT(1,0)}] - \frac{M}{4} E[e^{-sT(0,0)}] &= 0 \\
(M + s)E[e^{-sT(0,2)}] - \frac{M}{4} E[e^{-sT(2,0)}] - \frac{M}{4} E[e^{-sT(0,2)}] - \frac{M}{4} E[e^{-sT(0,0)}] &= 0 \\
-M E[e^{-sT(0,1)}] - \frac{M}{4} E[e^{-sT(0,0)}] - \frac{M}{4} E[e^{-sT(1,1)}] &= 0 \\
\end{align*}
\]

for $d_1 \in \mathbb{N}_0$ and $d_2 \in \mathbb{N}_0$.

As for the previous model, we calculate the moment-generating function of $T(d_1,d_2)$ using the theory of random walks rather than by solving these equations directly.

Denoting by $T^{(r)}_{(d_1,d_2)}$ the time until two individuals initially at "distance" $(d_1, d_2)$ are present in a single subpopulation for the first time,

\[ T(d_1,d_2) \overset{d}{=} T^{(r)}_{(d_1,d_2)} + T(0,0) \]

as before, with $T^{(r)}_{(d_1,d_2)}$ and $T(0,0)$ independent, so that

\[ E[e^{-sT(d_1,d_2)}] = E[e^{-sT^{(r)}_{(d_1,d_2)}}] \times E[e^{-sT(0,0)}]. \]

(3.44)

Until the two individuals are present in a single subpopulation for the first time, their distance (seen as a bivariate process with a component for each dimension of the lattice) performs a symmetric random walk on $\mathbb{N} \times \mathbb{N}$, where transitions occur at rate $M$. If $(d_1, d_2) \neq (0,0)$, the distribution of $T^{(r)}_{(d_1,d_2)}$ is that of the first passage time through $(0,0)$ of this random walk, starting at $(d_1, d_2)$. We denote by $(D_1, D_2) \equiv \{(D_1, D_2)_v : v = 0, 1, 2, \ldots \}$ the jump chain of this random walk ($(D_1, D_2)$ is a discrete-time symmetric random walk on $\mathbb{N} \times \mathbb{N}$). The probability that $(D_1, D_2)$, starting from $(d_1, d_2)$, is at the origin immediately after the $v$th step is given by Feller (1968), Chapter XIV, problem 28:

\[ P^{(v)}_{(d_1,d_2)(0,0)} := P\{(D_1, D_2)_v = (0,0) \mid (D_1, D_2)_0 = (d_1, d_2)\} \]

\[ = \frac{1}{(2\pi)^2} \int\int (\cos x + \cos y)^v \cos(d_1x) \cos(d_2y) \, dx \, dy \]

(3.45)

for $d_1, d_2 \in \mathbb{N}$. The generating function of the sequence $(P^{(v)}_{(d_1,d_2)(0,0)})_{v \in \mathbb{N}}$ is hence given by

\[ P_{(d_1,d_2)(0,0)}(z) := \sum_{v=0}^{\infty} P^{(v)}_{(d_1,d_2)(0,0)} z^v \]

\[ = \frac{1}{(2\pi)^2} \int\int \cos(d_1x) \cos(d_2y) \sum_{v=0}^{\infty} (\cos x + \cos y)^v \left( \frac{z}{2} \right)^v \, dx \, dy \]

\[ = \frac{1}{(2\pi)^2} \int\int \frac{\cos(d_1x) \cos(d_2y)}{1 - (\cos x + \cos y) z} \, dx \, dy \]

(3.46)
for $0 \leq z < 1$. As in the previous subsection (see equation (3.36)):

$$E[e^{-ST_{(d_1,d_2)}}] = \frac{P(d_1,d_2)(0,0) \left( \frac{M}{M+s} \right)}{P(0,0)(0,0) \left( \frac{M}{M+s} \right)}$$

(3.47)

for $d_1, d_2 \in N$. Substituting (3.46) into this equation, we find for $s > 0$:

$$E[e^{-ST_{(d_1,d_2)}}] = \frac{\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{\cos(x) \cos(y)}{M+s-M(\cos x+\cos y)^{1/2}} \, dx \, dy}{\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1}{M+s-M(\cos x+\cos y)^{1/2}} \, dx \, dy}$$

(3.48)

for $d_1, d_2 \in N$ and $s > 0$. In the context of the probability of identity by descent, Maruyama (1970) suggested a similar expression as an approximation valid under a stepping-stone model on a large square torus. The limit of Maruyama's result as the subpopulation size tends to infinity agrees with our value.

Equation (3.49) is in fact the limit of the corresponding result for the torus, equation (3.40), as the number of colonies in both dimensions of the torus tends to infinity. This convergence could easily have been proved directly from properties of the random walk performed by the distance between the two individuals before they enter a single subpopulation. Denoting the $\nu$-step transition probabilities from $(d_1, d_2)$ to $(0,0)$, introduced in equations (3.33) and (3.45), by $P^{(\nu)}(d_1, d_2)(0,0)(K,L)$ and $P^{(\nu)}(d_1, d_2)(0,0)(\infty, \infty)$, respectively, indicating whether they refer to the discrete-time symmetric random walk on the "$K \times L$ torus" \(\{0, \ldots, K-1\} \times \{0, \ldots, L-1\}\) or to that on the infinite two-dimensional lattice, we have for fixed $d_1, d_2, \nu \in N$:

$$P^{(\nu)}(d_1, d_2)(0,0)(K,L) = P^{(\nu)}(d_1, d_2)(0,0)(\infty, \infty) \text{ for } K > 2(d_1 + \nu) + 1 \text{ and } L > 2(d_2 + \nu) + 1,$$

because for such $K$ and $L$, the two components of the random walk cannot reach the positions \([K/2]\) or \([L/2]\), respectively, in $\nu$ steps. Hence

$$\lim_{K \to \infty} \lim_{L \to \infty} P^{(\nu)}(d_1, d_2)(0,0)(K,L) = P^{(\nu)}(d_1, d_2)(0,0)(\infty, \infty)$$

for every $d_1, d_2, \nu \in N$. By the dominated convergence theorem we also have convergence of the generating functions, for $0 \leq z < 1$:

$$\lim_{K \to \infty} \lim_{L \to \infty} \sum_{\nu=0}^{\infty} P^{(\nu)}(d_1, d_2)(0,0)(K,L) z^\nu = \sum_{\nu=0}^{\infty} P^{(\nu)}(d_1, d_2)(0,0)(\infty, \infty) z^\nu$$
for \( d_1, d_2 \in \mathbb{N} \). Because of equations (3.36) and (3.47) it follows that the first passage time through 
\((0,0)\) of the (continuous-time) random walk performed by the distance between two individuals
initially at distance \((d_1, d_2)\) on a \(K \times L\) torus converges in distribution to that of the random walk
described by the distance between two individuals initially at distance \((d_1, d_2)\) in the infinite two-
dimensional lattice, as \( K \) and \( L \) tend to infinity. Because of equations (3.39) and (3.48), it follows
that the coalescence time of two individuals under a stepping-stone model on a torus converges in
distribution, as the number of colonies in both dimensions of the torus tends to infinity, to that
of two individuals at the same distance on the infinite two-dimensional lattice.

The double integrals in result (3.49) can be reduced to single integrals, which are easier to
evaluate by computer using numerical integration. To do so, the results for the infinite linear
stepping-stone model prove very helpful. Equating the two different expressions found for the
moment-generating function of the coalescence time of two individuals \( d \) steps apart under the
infinite linear stepping-stone model, equations (3.27) and (3.28), we find the following identity,
valid for every \( d \in \mathbb{N} \) and for \( s > 0 \):

\[
\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\cos(xd)}{M + s - M \cos x} \, dx = \frac{(M + s - \sqrt{(2M + s)s})^d}{M^d \sqrt{(2M + s)s}}. \tag{3.50}
\]

Using this identity with \( s + \frac{M}{2}(1 - \cos x) \) instead of \( s \), and with \( \frac{M}{2} \) instead of \( M \), we can re-write
the double integral in the numerator of (3.49) as

\[
\frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{\cos(d_1x) \cos(d_2y)}{M + s - M(\cos x + \cos y)/2} \, dx \, dy
\]

\[
= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(d_1x) \left[ \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\cos(d_2y)}{M + s - \frac{M}{2}(1 - \cos x) - \frac{M}{2} \cos y} \, dy \right] \, dx
\]

\[
= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(d_1x) \left( \frac{M + s - \frac{M}{2} \cos x - \sqrt{(M + s - \frac{M}{2} \cos x)^2 - (\frac{M}{2})^2}}{\sqrt{(M + s - \frac{M}{2} \cos x)^2 - (\frac{M}{2})^2}} \right)^d \, dx
\]

\[
= \frac{2}{\pi M} \int_{0}^{\pi} \cos(d_1x) \left( \frac{2 + 2s/M - \cos x - \sqrt{(2 + 2s/M - \cos x)^2 - 1}}{\sqrt{(2 + 2s/M - \cos x)^2 - 1}} \right)^d \, dx
\]

so that

\[
E[e^{-\tau(d_1,d_2)}] = \frac{2 \int_{0}^{\pi} \cos(d_1x) \left( \frac{2 + 2s/M - \cos x - \sqrt{(2 + 2s/M - \cos x)^2 - 1}}{\sqrt{(2 + 2s/M - \cos x)^2 - 1}} \right)^d \, dx}{\pi M + 2 \int_{0}^{\pi} \frac{1}{\sqrt{(2 + 2s/M - \cos x)^2 - 1}} \, dx}. \tag{3.51}
\]

for \( d_1, d_2 \in \mathbb{N} \) and \( s > 0 \). Malécot (1950, 1975) obtained a similar expression for the probability
of identity by descent of a pair of genes under a more general stepping-stone model than the
one described in this paper. It can be shown that under our model, Malécot's result gives an
approximation of result (3.51) (with little actual simplification) by focusing on the value of the
integrand in the neighbourhood of the singularity \( x = 0 \), valid when both \( s/M \) and \( d_2 s/M \) are
negligible relative to 1.
Because the unbounded two-dimensional symmetric random walk is null-recurrent, the mean and the second moment of the coalescence time of any two individuals are infinite:

\[
ET_{d_1,d_2} = \infty \\
E[T_{d_1,d_2}^2] = \infty.
\]

(3.52)  (3.53)

3.3 Some less symmetric structures

The symmetric island model and the stepping-stone models studied in the previous sections are the models traditionally dealt with in the literature. The structure of natural populations is, of course, far less symmetric. In particular, most real populations have subpopulations of unequal sizes. However, little effort has been made to investigate what effect asymmetries in the population structure may have on both the genealogy and the genetic composition of the population. In this section we study three particular models of population structure which allow for unequal subpopulation sizes, different migration patterns from different subpopulations or different migration rates between different pairs of subpopulations. In the first subsection we consider a general two-population model, in which the population consists of two subpopulations of possibly different sizes. In Subsection 2 we introduce the “continental island model”, where one subpopulation (the “continent”) has a migration pattern different from that of the other subpopulations. Subsection 3 focuses on a four-population structure with unequal migration rates.

3.3.1 The general two-population model

![Figure 3.6: A general two-population model with $c_1 > c_2$.](image)

The population is divided into two subpopulations ($n = 2$) of possibly different sizes. Every generation, a fixed number of individuals is exchanged between the two subpopulations.

We assume the model of Section 2.1. In the notation of that model, $S = \{1, 2\}$. Subpopulation 1 contains $2c_1N$ haploid individuals, while the size of subpopulation 2 is $2c_2N$. Reproduction in each subpopulation follows the neutral Wright-Fisher model, introduced in Section 2.1. Every generation, a proportion $q_{12}$ of subpopulation 1 is exchanged for a proportion $q_{21}$ of subpopulation 2, where

\[
c_1 q_{12} = c_2 q_{21}
\]

(3.54)
Chapter 3: GENEALOGY

(see Assumption 2.1) so that the number of migrants in both directions is the same and the subpopulation sizes remain constant under migration. Under the assumptions of Theorem 2.1 and measuring time in units of $2N$ generations, the genealogy of a sample from this population is well described by the structured coalescent, whose $Q$-matrix is given by equation (2.2), where for $i,j \in \{1,2\}$ with $j \neq i$:

$$M_i = M_{ij} = \lim_{N \to \infty} \left(4N \frac{c_j}{c_i} q_{ij}\right) = \lim_{N \to \infty} (4N q_{ij}),$$

(see the definition of $M_{ij}$ in Theorem 2.1 and see equation (2.1)), where the last equality follows from (3.54). Working backward in time, each individual in subpopulation $i$ has scaled rate $M_i/2$ of migrating to the other subpopulation ($i = 1,2$). Because the scaled migration rates $M_1$ and $M_2$ are in part determined by the subpopulation sizes and because in Chapter 4 we shall compare results for different relative subpopulation sizes, we express all results in terms of the absolute quantity

$$M := c_1 M_1 = c_2 M_2$$

(see equation (3.54)), which is twice the number of individuals exchanged between the two subpopulations every generation. In this context, the term "migration rate" will refer to $M$. For $c_1 = c_2$, the two-population model is, up to a factor of time-scaling, the symmetric island model described in Section 1, with $n = 2$ colonies.

Denoting by $T_{ij}$ the coalescence time of an individual in subpopulation $i$ and an individual in subpopulation $j$ ($i,j = 1,2$), Theorem 2.4 gives the following equations for the moment-generating function of the coalescence time a pair of individuals from this population:

$$\left\{ \begin{align*}
&\left(\frac{1}{c_i} + M_1 + s\right) E[e^{-s T_{11}}] - M_1 E[e^{-s T_{12}}] = \frac{1}{c_i} \\
&\left(\frac{1}{c_2} + M_2 + s\right) E[e^{-s T_{21}}] - M_2 E[e^{-s T_{22}}] = \frac{1}{c_2} \\
&\left(\frac{1}{c_1} + M_1 + s\right) E[e^{-s T_{12}}] - M_1 E[e^{-s T_{12}}] - M_2 E[e^{-s T_{12}}] = 0
\end{align*} \right\}
$$

We recall the notation $c = c_1 + c_2$ (making the total population size $2cN$) and we denote by $P := c_1/c$ the proportion of the total population that lives in subpopulation 1. Expressed in terms of $P$, $c$ and the "migration rate" $M$, the solution of this system of linear equations is given by

$$\begin{align*}
E[e^{-s T_{11}}] &= \frac{M + 2P(1-P)cs}{A} \{1 + M + (1-P)cs\} \tag{3.56} \\
E[e^{-s T_{12}}] &= \frac{M}{A} \left\{ 1 + M + 2P(1-P)cs \right\} \tag{3.57} \\
E[e^{-s T_{22}}] &= \frac{M + 2P(1-P)cs}{A} \{1 + M + Pcs\} \tag{3.58}
\end{align*}$$

where

$$A = P(1-P) \left\{ Mcs(4 + 3cs) + 2cs(1 + cs) + 2P(1-P)(cs)^2 \right\} + M(1 + M)(1 + cs).$$
The mean and the second moment of the coalescence time of a pair of individuals are obtained either by differentiating these results for the moment-generating function with respect to $s$, once and twice, respectively, and taking the limit as $s \downarrow 0$, or by solving the equations given by Theorems 2.5 and 2.6. We find:

$$ET_{11} = c - \frac{(1-P)(1-2P)c}{1+M}$$

(3.59)

$$ET_{12} = c + \frac{2P(1-P)c}{M}$$

(3.60)

$$ET_{22} = c + \frac{P(1-2P)c}{1+M}$$

(3.61)

and

$$E[T^2_{11}] = 2c^2 + 2(1-P)c^2 \left( \frac{M(3P-1)}{(1+M)^2} + \frac{P-1+10P^2-8P^3}{(1+M)^2} + \frac{4P^2(1-P)}{M(1+M)^2} \right)$$

(3.62)

$$E[T^2_{12}] = 2c^2 + 2P(1-P)c^2 \frac{4P(1-P)(1+2M) + M(2+3M)}{M^2(1+M)}$$

(3.63)

$$E[T^2_{22}] = 2c^2 + 2Pc^2 \left( \frac{M(2-3P)}{(1+M)^2} + \frac{2+3P-14P^2+8P^3}{(1+M)^2} + \frac{4P(1-P)^2}{M(1+M)^2} \right).$$

(3.64)

In the previous sections we observed that in the symmetric island model and in the stepping-stone models on the circle and the torus, the mean coalescence time of two individuals from a single subpopulation is independent of the migration rate and equal to that of two individuals from a panmictic population of the same total size. Note that this is not the case under the general two-population model, unless the two colonies have the same size ($P = 1/2$). Results (3.59) and (3.61) show that under this model, the mean coalescence time of a pair of individuals from the larger colony is larger than $c$ (which is the mean coalescence time of two individuals from a panmictic population of the same total size and with the same time-scaling), while the mean coalescence time of two individuals from the smaller subpopulation is smaller than in the panmictic case. Both $ET_{11}$ and $ET_{22}$ converge to $c$ as the migration rate $M$ tends to infinity. In Chapter 5 we will study under what conditions the mean coalescence time of a pair of individuals from a single subpopulation is independent of the migration rate and equal to that of two individuals from a panmictic population of the same total size.

### 3.3.2 The continental island model

The population is divided into a central subpopulation (the "continent") surrounded by a number of "islands". Migration occurs only between the continent and each island. Individuals cannot directly (that is, in one generation) migrate from one island to another but have to go via the continent. We assume all the islands are identical with respect to size and with respect to migration rate to the continent.

We label the continent as subpopulation 0 and the islands as subpopulations 1 to $n-1$ ($n$ is the number of subpopulations, including the continent); $S = \{0, 1, \ldots, n-1\}$. The size of the continent
is $2c_0 N$, while the islands each contain $2c_1 N$ haploid individuals. Every generation an equal number of individuals are exchanged between the continent and each of the islands. Reproduction and migration are assumed to follow the model of Section 2.1. In the notations of that model, $q_{01}$ is the proportion of the continent migrating to each of the islands, every generation, and $q_{10}$ is the proportion of each island that migrates to the continent every generation (for $i = 1, \ldots, n - 1$: $q_{0i} = q_{01}$ and $q_{1i} = q_{10}$). The requirement that migration does not affect the subpopulation sizes, Assumption 2.1, gives

$$c_0 q_{01} = c_1 q_{10}. \quad (3.65)$$

Measuring time in units of $2N$ generations, the genealogy of a sample from this population is (under the assumptions of Theorem 2.1) well described by the structured coalescent. In equation (2.2) for its Q-matrix, the scaled migration rates are given by

$$M_{0i} = \lim_{N \to \infty} (4Nq_{01}) \quad \text{for } i = 1, \ldots, n - 1$$
$$M_{1i} = \lim_{N \to \infty} (4Nq_{1i}) \quad \text{for } i = 1, \ldots, n - 1$$
$$M_{ij} = 0 \quad \text{for } i, j = 1, \ldots, n - 1 \text{ with } j \neq i,$$

where we have combined the definition of $M_{ij}$ (see Theorem 2.1), equation (2.1) and equation (3.65). The total scaled migration rate from the continent is $M_0 = (n - 1)M_{01}$, while that from each island is $M_1 = M_{10}$. Working backward in time, each individual in the continent has scaled rate $M_0/(2(n - 1))$ of migrating to any particular island, while every individual in the islands has scaled rate $M_1/2$ of migrating to the continent.

We denote by $T_{ij}$ the coalescence time of an individual from subpopulation $i$ and an individual from subpopulation $j$. Because all islands are equivalent, there are essentially four possibilities for
the locations of a pair of individuals from the population. So

\[ T_{ii} \overset{d}{=} T_{1i} \quad \text{for } i = 1, \ldots, n - 1 \]
\[ T_{0i} \overset{d}{=} T_{01} \quad \text{for } i = 1, \ldots, n - 1 \]
\[ T_{ij} \overset{d}{=} T_{12} \quad \text{for } i, j = 1, \ldots, n - 1 \text{ with } j \neq i, \]

and the equations of Theorem 2.4 reduce for a sample of two individuals to

\[
\begin{align*}
\left( \frac{1}{c_0} + M_0 + s \right) E[e^{-sT_{00}}] - M_0 E[e^{-sT_{10}}] &= \frac{1}{c_0} \\
\left( \frac{1}{c_1} + M_1 + s \right) E[e^{-sT_{11}}] - M_1 E[e^{-sT_{10}}] &= \frac{1}{c_1} \\
\left( \frac{M_2}{2} + M_1 + s \right) E[e^{-sT_{01}}] - \frac{M_2}{2} E[e^{-sT_{11}}] - \frac{M_0}{2(n-1)} E[e^{-sT_{12}}] &= -\frac{M_0}{2(n-1)} E[e^{-sT_{12}}] = 0 \\
(M_1 + s) E[e^{-sT_{12}}] - M_1 E[e^{-sT_{01}}] &= 0,
\end{align*}
\]

where the last equation and the last term in the left-hand side of the third equation are present only if there are at least two islands \((n \geq 3)\). In order to compare results for different relative sizes of continent and islands, which we will do in Chapter 4, it is convenient to define the "migration rate"

\[ M := c_1 M_1 = c_0 M_0 / (n - 1) \quad (3.66)\]

(see equation (3.65)), which is twice the number of individuals exchanged between the continent and each of the islands every generation. Expressed in terms of the parameters \(M, n, c = c_0 + (n - 1)c_1\) (making the total population size \(2cN\)) and the proportion \(P := c_0/c\) of individuals in the continent, the solution of this system of linear equations gives for \(i, j = 1, \ldots, n - 1\) with \(j \neq i:\)

\[
\begin{align*}
E[e^{-sT_{00}}] &= \frac{B + 2(1 - P^2)(n-1)^2M^2cs + (1 + 4P)(1 - P)^2(n-1)M(cs)^2 + 2P(1 - P)^2(cs)^3}{D} \quad (3.67) \\
E[e^{-sT_{11}}] &= \frac{B + (n - 1 + 2P - 2P^2)(n-1)^2M^2cs + 3P(1 - P)(n-1)^2M(cs)^2 + 2(n-1)P^2(1 - P)^2(cs)^3}{D} \quad (3.68) \\
E[e^{-sT_{01}}] &= \frac{(n-1)M \{(n-1)M + (1 - P)cs\} + (n-1)M + (n-2)P + 2P(1 - P)cs}{D} \quad (3.69) \\
E[e^{-sT_{12}}] &= \frac{(n-1)^2M^2 \{(1 + (n-1)M + (n-2)P + 2P(1 - P)cs\}}{D} \quad (3.70)
\end{align*}
\]

where

\[
\begin{align*}
B &= (n-1)^3M^3 + (1 - 2P + nP)(n-1)^2M^2 + (1 - P)(1 + 2P)(n-1)^2Mcs \\
&\quad + 2(n-1)P(1 - P)^2(cs)^2, \\
D &= B + \{(n-1)M + n + 3P - 4P^2\} (n-1)^2M^2cs \\
&\quad + (1 - P) \{1 + 3Pn - 4P^2 + (1 + 3P)(n-1)M\} (n-1)M(cs)^2 \\
&\quad + P(1 - P)^2 \{2 - 4P + 2P + (3 + 2P)(n-1)M\} (cs)^3 + 2P^2(1 - P)^3(cs)^4.
\end{align*}
\]

For \(n = 2\), results (3.67) to (3.69) reduce to the results found in the previous subsection.
Chapter 3: GENEALOGY

The mean and the second moment of the coalescence time of a pair of individuals are obtained from the equations of Theorems (2.5) and (2.6) or by differentiation of the results for the moment-generating function. We find for $i, j = 1, \ldots, n - 1$ with $j \neq i$:

\[
ET_{ii} = c - \frac{c(1-P)(2P+n-3)}{1+(n-1)M+(n-2)P} \quad (3.71)
\]

\[
ET_{ij} = c + \frac{c(1-P)(2P+n-2)}{1+(n-1)M+(n-2)P} + \frac{c(1-P)(nP+n-2)}{(n-1)M\{1+(n-1)M+(n-2)P\}} \quad (3.72)
\]

and

\[
ET_{00} = c + \frac{c(1-P)(2P+n-3)}{1+(n-1)M+(n-2)P} \quad (3.73)
\]

\[
ET_{0i} = c + \frac{c(1-P)(2P+n-1)}{1+(n-1)M+(n-2)P} + \frac{c(1-P)(2P+1)}{M\{1+(n-1)M+(n-2)P\}} \quad (3.74)
\]

\[
E[T^2_{00}] = 2c^2 + 2c^2(1-P) \left[ \frac{2n-5+3P}{(n-1)M} \right] \quad (3.75)
\]

\[
E[T^2_{ii}] = 2c^2 \left[ \frac{n-2-2Pn+6P-3P^2}{(n-1)M} \right] \quad (3.76)
\]

\[
E[T^2_{ij}] = 2c^2 + 2c^2(1-P) \left[ \frac{n-2+3P}{(n-1)M} \right] \quad (3.77)
\]

\[
E[T^2_{0i}] = 2c^2 + 2c^2(1-P) \left[ \frac{2n-3(1-P)}{(n-1)M} \right] \quad (3.78)
\]

Note that again, the mean coalescence time of two individuals from the continent, result (3.71),
and that of two individuals from a single island, result (3.72), both depend on the migration rate, whatever the relative sizes of continent and islands may be (unless \( n = 2 \) and \( P = 1/2 \), which is a symmetric island model with two colonies). If there is more than one island \((n > 2)\), the mean coalescence time of a pair of individuals from the continent is always larger than \( c \), while that of two individuals from the same island is always smaller than \( c \), even if the continent is smaller than the islands (\( c \) is the mean coalescence time of two individuals from a panmictic population of the same total size and with the same time-scaling as the subdivided population). Note that \( \lim_{M \to \infty} ET_{ii} = c \) for \( i = 0, 1, \ldots, n - 1 \).

### 3.3.3 Four-population model with geographic barrier

![Four-population model diagram](image)

Figure 3.8: The four-population model with geographic barrier.

The population consists of four equal-sized subpopulations, with a geographic barrier (such as a river or a mountain) separating colonies 1 and 2 from colonies 3 and 4. Every generation, a proportion \( m \) of each subpopulation is replaced by immigrants, of which a fraction \((1 - \epsilon)\) come from the other subpopulation at the same side of the barrier and a fraction \( \epsilon \) come from the "neighbouring" subpopulation at the other side of the barrier \((0 < \epsilon < 1)\). A value of \( \epsilon \) close to zero indicates there is little gene flow across the barrier, while \( \epsilon = 1/2 \) yields the circular stepping-stone model with four colonies (no barrier). A value of \( \epsilon \) larger than \( 1/2 \) corresponds to a barrier in the other direction, separating subpopulations 1 and 4 from subpopulations 2 and 3.

Under the appropriate assumptions and time-scale, for example those of Section 2.1 and Theorem 2.1, the genealogy of a sample from this population is well described by the structured coalescent, with in equation (2.2) for its Q-matrix:

\[
S = \{1, 2, 3, 4\}
\]

\[
c_i = 1 \text{ for } i = 1, 2, 3, 4
\]

and the scaled migration rates

\[
M_{12} = M_{34} = (1 - \epsilon)M
\]
\[ M_{14} = M_{23} = \epsilon M \]

\[ M_{13} = M_{24} = 0, \]

while for \( i, j = 1, 2, 3, 4 \): \( M_{ij} = M_{ji} \). If the Wright-Fisher model is assumed for reproduction within subpopulations and each subpopulation contains \( 2N \) haploid individuals, time-scaling is in units of \( 2N \) generations and \( M = \lim_{N \to \infty} (4Nm) \), which is twice the number of immigrants per subpopulation per generation. In the notation of Section 2.1 we have for the migration rates in discrete time: \( q_{12} = q_{24} = (1 - \epsilon)m \), \( q_{14} = q_{23} = \epsilon m \), \( q_{13} = q_{24} = 0 \) and \( q_{11} = q_{22} \).

We denote by \( T_{ij} \) the coalescence time of an individual from subpopulation \( i \) and an individual from subpopulation \( j \). Because all four subpopulations (though not pairs of subpopulations) are identical with respect to size, migration pattern and migration rates, there are essentially four possibilities for the locations of a sample of two individuals from this population. So

\[
\begin{align*}
T_{11} & \overset{d}{=} T_{22} \overset{d}{=} T_{33} \overset{d}{=} T_{44}, \\
T_{12} & \overset{d}{=} T_{34}, \\
T_{13} & \overset{d}{=} T_{24}, \\
T_{14} & \overset{d}{=} T_{23}.
\end{align*}
\]

The system of equations given by Theorem 2.4 for the moment-generating function of the coalescence time of a pair of individuals thus reduces to:

\[
\begin{align*}
(1 + M + s)E[e^{-sT_{11}}] - (1 - \epsilon)ME[e^{-sT_{12}}] - \epsilon ME[e^{-sT_{14}}] & = 1 \\
(M + s)E[e^{-sT_{12}}] - (1 - \epsilon)ME[e^{-sT_{11}}] - \epsilon ME[e^{-sT_{13}}] & = 0 \\
(M + s)E[e^{-sT_{13}}] - \epsilon ME[e^{-sT_{12}}] - (1 - \epsilon)ME[e^{-sT_{14}}] & = 0 \\
(M + s)E[e^{-sT_{14}}] - \epsilon ME[e^{-sT_{13}}] - (1 - \epsilon)ME[e^{-sT_{12}}] & = 0.
\end{align*}
\]

The solution of this system of linear equations yields

\[
\begin{align*}
E[e^{-sT_{11}}] & = E[e^{-sT_{22}}] = E[e^{-sT_{33}}] = E[e^{-sT_{44}}] = \frac{(M + s)s^2 + 2Ms + 2(1 - \epsilon)M^2}{G} \\
E[e^{-sT_{12}}] & = E[e^{-sT_{34}}] = \frac{(1 - \epsilon)M\left(s^2 + 2Ms + 2(1 - \epsilon)M^2\right)}{G} \\
E[e^{-sT_{13}}] & = E[e^{-sT_{24}}] = \frac{2\epsilon(1 - \epsilon)(M + s)M^2}{G} \\
E[e^{-sT_{14}}] & = E[e^{-sT_{23}}] = \frac{\epsilon M\left(s^2 + 2Ms + 2(1 - \epsilon)M^2\right)}{G}
\end{align*}
\]

where

\[
G = s^4 + 4Ms^3 + 4\{1 + \epsilon(1 - \epsilon)\}M^2s^2 + 8\epsilon(1 - \epsilon)M^3s \\
+ (M + s)\{s^2 + 2Ms + 2\epsilon(1 - \epsilon)M^2\}.
\]
Maruyama (1970) calculated the probability of identity by descent of a pair of genes under a stepping-stone model on a $K \times K$ torus with different migration rates in the two dimensions of the torus, in the case of finite subpopulation size. Taking the limit as the subpopulation size tends to infinity, Maruyama's result simplifies for $K = 2$ to equations (3.79) to (3.82).

The mean and the second moment of the coalescence time of a pair of individuals are found by differentiating the results for the moment-generating function or by solving the equations given by Theorems 2.5 and 2.6. We find:

\[
\begin{align*}
ET_{11} &= ET_{22} = ET_{33} = ET_{44} = 4 \\
ET_{12} &= ET_{34} = 4 + \frac{1}{M} + \frac{1}{(1 - \epsilon)M} \\
ET_{13} &= ET_{24} = 4 + \frac{1}{\epsilon(1 - \epsilon)M} \\
ET_{14} &= ET_{23} = 4 + \frac{1}{M} + \frac{1}{\epsilon M}
\end{align*}
\]

and

\[
\begin{align*}
E[T_{11}^2] &= E[T_{22}^2] = E[T_{33}^2] = E[T_{44}^2] = 32 + \frac{4}{M} + \frac{4}{\epsilon(1 - \epsilon)M} \\
E[T_{12}^2] &= E[T_{34}^2] = 32 + \frac{12}{M} + \frac{2}{M^2} + \frac{8}{(1 - \epsilon)M} + \frac{1}{(1 - \epsilon)M^2} + \frac{4}{\epsilon(1 - \epsilon)M} + \frac{2}{\epsilon(1 - \epsilon)^2 M^2} \\
E[T_{13}^2] &= E[T_{24}^2] = 32 + \frac{4}{M} + \frac{12}{\epsilon(1 - \epsilon)M} - \frac{1}{\epsilon(1 - \epsilon)M^2} + \frac{2}{\epsilon^2(1 - \epsilon)^2 M^2} \\
E[T_{14}^2] &= E[T_{23}^2] = 32 + \frac{12}{M} + \frac{2}{M^2} + \frac{8}{\epsilon M} + \frac{1}{\epsilon M^2} + \frac{4}{\epsilon(1 - \epsilon)M} + \frac{2}{\epsilon^2(1 - \epsilon)^2 M^2}.
\end{align*}
\]

Note that the mean coalescence time of two individuals from the same colony under this model, result (3.83), does not depend on the migration rate and is equal to $c = \sum_{i=1}^{4} c_i = 4$. In Chapter 5 this will be seen to be a consequence of the fact that all subpopulations are identical with respect to size, migration pattern and migration rates, as is also the case under the symmetric island and stepping-stone models studied in the previous sections.
Chapter 4

SUBPOPULATION DIFFERENTIATION UNDER VARIOUS MODELS OF POPULATION STRUCTURE

Whereas the previous chapters were concerned with genealogy, we will in this chapter focus on genetic variation, maintained by neutral mutation. The genealogical results obtained in Chapter 3 are used to study the genetic differentiation among the subpopulations of a subdivided population. Measured by Wright's coefficient $F_{ST}$ (to be defined below), the amount of subpopulation differentiation can be expressed in terms of the respective coalescence times of pairs of genes sampled within and among subpopulations. For the range of models of population structure for which we calculated the moment-generating function, mean and second moment of the coalescence time of a pair of genes in the previous chapter, we will now calculate the value of $F_{ST}$ explicitly. The dependence of the amount of subpopulation differentiation (as measured by $F_{ST}$) on the parameters of population structure and on the neutral mutation rate will be examined and compared between the different population structures considered. We show that, contrary to common belief, the dependence of $F_{ST}$ on the mutation rate can be very strong, even at very small mutation rates, particularly when the population occupies an essentially one-dimensional habitat and the number of subpopulations is large.

In the first section of this chapter we introduce Wright's coefficient $F_{ST}$, which serves as a measure of the genetic differentiation among the subpopulations of a subdivided population (Subsection 1.1). In Subsection 1.2, $F_{ST}$ is expressed in exact and approximate ways in terms of
coalescence times of pairs of genes. In Section 2 we calculate the values of $F_{ST}$ and its approximations under the symmetric island and stepping-stone models studied in the previous chapter. For each model, figures illustrate how $F_{ST}$ depends on the various parameters of the model. Particular attention is paid to the dependence of $F_{ST}$ on the mutation rate, which is compared between the different population structures. Section 3 is concerned with the effect which asymmetries in the population structure (for example, unequal subpopulation sizes) may have on the amount of subpopulation differentiation, as measured by $F_{ST}$. The values of $F_{ST}$ and its approximations are calculated for the three “less symmetric” models of population structure introduced in the previous chapter. Figures show for each model how $F_{ST}$ depends on the different parameters of the model and, in particular, on the level of asymmetry in the structure.

4.1 Wright’s coefficient $F_{ST}$

4.1.1 Introduction

The effect of population structure on the genetic composition of a population has traditionally been analyzed in terms of Wright’s hierarchical $F$-statistics, which are essentially inbreeding coefficients. Inbreeding in a diploid population occurs when mates tend to be more closely related than they would be if the population were randomly mating. Its effect is to decrease the amount of heterozygosity (that is, the proportion of diploid individuals whose two genes at a particular locus carry different alleles), compared to the amount of heterozygosity one would expect from the allele frequencies if the population were under Hardy-Weinberg equilibrium. Population subdivision, like inbreeding, has a reduction of heterozygosity for effect. Studying a population subdivided into distinct colonies, Wright (1951) separated the respective contributions towards inbreeding of non-random mating within colonies and of the population subdivision itself. The inbreeding-like effect of the population subdivision is represented by Wright’s $F_{ST}$, originally defined as the correlation coefficient (relative to the total population) between two alleles drawn at random from the same subpopulation. Under the assumptions that are implicit in Wright (1951), this correlation coefficient is equal to the relative decrease in heterozygosity that is caused by the population subdivision, i.e.

$$F_{ST} = \frac{H_T - H_S}{H_T},$$  \hspace{1cm} (4.1)

where $H_T$ and $H_S$ are the heterozygosities one would expect from the allele frequencies under Hardy-Weinberg equilibrium in, respectively, the total population and each subpopulation. Equa-

\footnote{Under Hardy-Weinberg equilibrium, the heterozygosity expected at a particular locus is $\sum_{i=1}^{J} p_i(1 - p_i)$, where $J$ is the number of alleles in the population at the locus under consideration and $p_i$ is the frequency of the $i$th allele ($i = 1, \ldots, J$). Hardy-Weinberg equilibrium corresponds to the situation of neutral genes in a randomly mating population (under conditions set out, for example, in Hartl 1988).}
tion (4.1) now widely serves as the definition of $F_{ST}$ (Nei 1973, Hartl 1988), although other definitions are also in use (see for example Chakraborty and Danker-Hopfe 1991 for a review).

The coefficient $F_{ST}$ has commonly been used as a measure of the genetic differentiation among subpopulations. Its major application is to estimating the level of gene flow between the subpopulations (Slatkin 1985 and references therein, Slatkin and Barton 1989, Slatkin 1991, 1993). Denoting by $N$ the number of diploid individuals per subpopulation and by $m$ the fraction of each subpopulation that is replaced by immigrants every generation, the level of gene flow $Nm$ is estimated using the formula

$$Nm \approx \frac{1}{4} \left( \frac{1}{F_{ST}} - 1 \right)$$

(Wright 1931), which is based on the symmetric island model of population structure, described in Section 3.1, and on the neutral Wright-Fisher model of reproduction. Estimates of $Nm$ obtained for pairs of subpopulations can be used to detect isolation by distance in a natural population and to test specific hypotheses about the structure and history of the population (Slatkin 1993). As equation (4.2) assumes a symmetric island model with a large number of subpopulations and a small mutation rate, it is important to understand how $F_{ST}$ depends on the real structure of the population and on the mutation rate.

Slatkin (1991) brought the relationship between $F_{ST}$ and genealogy to the foreground and introduced an approximation for $F_{ST}$, valid for small mutation rates, in terms of the mean time since the most recent common ancestor of two genes sampled from a single subpopulation, and that of two genes sampled from the total population. Calculating these mean coalescence times, he obtained (approximate) analytical $F_{ST}$ values for the symmetric island model and for the stepping-stone models on the circle and the torus. Using the genealogical results obtained in Chapter 3, we can calculate the exact\(^2\) values of $F_{ST}$ under the various models of population structure considered there. These exact $F_{ST}$ values make it possible to investigate how $F_{ST}$ depends on the mutation rate in different structured populations, and to examine the accuracy of Slatkin’s approximation for $F_{ST}$ (which is the limit of $F_{ST}$ as the mutation rate tends to zero). We will also investigate an alternative approximation for $F_{ST}$.

In accordance with the intrinsic assumptions of the structured coalescent, the analysis in this chapter is for selectively neutral genes at a single locus, subject to mutation and migration but without intragenic recombination. Our study is exact only for haploid species, and for diploid species with exclusively gametic migration. Nagylaki (1983) and Sawyer (1976) have set out conditions under which models of truly diploid migration are well approximated by the model of gamete migration.

\(^2\)Under the infinite alleles model for mutation (see the next subsection) and within the diffusion time-scale approximation, which is inherent in the coalescent approach.
4.1.2 $F_{ST}$ in terms of coalescence times

As in the previous chapters, we consider a population divided into partially isolated subpopulations, which are all large and panmictic, and we assume that the structured coalescent is an appropriate description of genealogy. The coefficient $F_{ST}$ can be related to the respective coalescence times of pairs of genes sampled within and among subpopulations. Slatkin (1991) did this in an approximate way, and in fact it can be done exactly. The distribution of these coalescence times can be found from the structured coalescent with $n_0 = 2$.

The exact $F_{ST}$ value: The coefficient $F_{ST}$ has been expressed in terms of probabilities of identity (Wright 1951, Nei 1973, Slatkin 1985, 1991) as:

$$ F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}}, $$

where $f_0$ is the probability that two genes sampled at random from a single subpopulation are identical (that is, carry the same allele), and $\bar{f}$ is the probability that two genes randomly chosen from the collection of subpopulations considered are identical. We note that equation (4.3), which is the definition of $F_{ST}$ adopted throughout this thesis, is appropriate for haploid as well as diploid populations. As is implicit in Slatkin and Barton (1989) and Slatkin (1985, 1991, 1993), we will approximate the probability of identity of two genes by the probability that since their descent from a common ancestor, neither gene has undergone a mutation (in which case the two genes are said to be "identical by descent"). This approximation is exact for the infinite alleles model of mutation, in which every mutant gene is assumed to be of a novel type.

Sampling two genes at random from the same subpopulation, the distribution of the time $T_0$ since their most recent common ancestor follows from the structured coalescent. Given $T_0$, the probability of identity of the two genes is, under the infinite alleles assumption, the probability that no mutation has occurred on either gene's lineage during time $T_0$. In the coalescent approximation, the probability that a particular gene has not mutated during time $T_0$ is $e^{-\theta T_0}$, where $\theta$ is the scaled mutation rate. For example, if we assume that each subpopulation contains $2N$ genes and evolves according to the neutral Wright-Fisher model, time-scaling is in units of $2N$ generations and $\theta = \lim_{N \to \infty} (4N \mu)$, where $\mu$ is the probability of mutation per gene per generation. Assuming that different genes mutate independently, it follows that

$$ f_0 = E[e^{-\theta T_0}] $$

In this context, random sampling is to be interpreted as follows: first choose a subpopulation according to some (specified) probability distribution on the subpopulations, and then choose a gene uniformly at random from the chosen subpopulation. In the case of equal-sized subpopulations (Section 2), the distribution on the subpopulations is assumed to be uniform. For the models with unequal-sized subpopulations, we will consider two different distributions on the subpopulations, indicated as "equal weighting" and "weighting by size" (see Section 3).

4The collection of subpopulations considered is in general either the total population (as is usually the case for island models) or a pair of subpopulations a specified distance apart (as is common for stepping-stone models).
Chapter 4: SUBPOPULATION DIFFERENTIATION

(Hudson 1990). Similarly

\[ f = E[e^{-\theta T}] \] (4.5)

where \( T \) is the coalescence time of two genes randomly sampled from the collection of subpopulations considered. The distribution of \( T \) also follows from the structured coalescent.

Note that (4.4) and (4.5) are the moment-generating functions of \( T_0 \) and \( T \), respectively, evaluated in \( \theta \). So for the infinite alleles model, \( F_{ST} \) can be written in an exact way in terms of the moment-generating functions of the coalescence times of different pairs of genes:

\[ F_{ST} = \frac{E[e^{-\theta T_0}] - E[e^{-\theta T}]}{1 - E[e^{-\theta T}]} \] (4.6)

Slatkin's approximation \( F_{ST}^{(0)} \): Slatkin (1991) suggested that \( F_{ST} \) might be approximated by its limit as the mutation rate tends to zero, which, by applying l'Hôpital's rule, can be expressed in terms of mean coalescence times as

\[ F_{ST}^{(0)} := \lim_{\theta \to 0} F_{ST} = \frac{ET - ET_0}{ET} \] (4.7)

provided these mean coalescence times are finite.

One advantage of Slatkin's approximation is the relative simplicity of both the calculation of the mean coalescence times and the resulting expression for \( F_{ST}^{(0)} \). This is largely due to the fact that for sufficiently symmetric models of population structure the mean coalescence time of two genes from a single colony \( (ET_0) \) does not depend on the migration rates or pattern, but depends only on the total population size and the reproductive model assumed (Strobeck 1987; see also Chapter 5). Another argument supporting the use of \( F_{ST}^{(0)} \) is the fact that it is, of course, independent of the mutation rate, which is helpful since in most cases, mutation rates are unknown, but known to be very small. Furthermore, it is commonly believed that for small mutation rates, \( F_{ST} \) is nearly independent of the mutation rate (Crow and Aoki 1984, Slatkin and Barton 1989, Slatkin 1985, 1991, 1993). However, as we will show in the next sections, the latter is not generally true. Nevertheless, Slatkin's approximation turns out to be accurate under most realistic conditions.

The first order approximation \( F_{ST}^{(1)} \): By considering one more term in the Taylor expansion of \( F_{ST} \) about mutation rate zero, one obtains an approximation for \( F_{ST} \) which is linear in the mutation rate and which, for mutation rates close to zero, should be more accurate than \( F_{ST}^{(0)} \):

\[ F_{ST}^{(1)} := \lim_{\theta \to 0} F_{ST} + \theta \lim_{\theta \to 0} \frac{\partial}{\partial \theta} F_{ST}. \] (4.9)

By repeatedly applying l'Hôpital's rule, this first order approximation can be expressed in terms of first and second moments (or mean and variance) of coalescence times of pairs of genes, as

\[ F_{ST}^{(1)} = \frac{ET - ET_0}{ET} - \theta \frac{E[T^2] ET_0 - E[T_0^2]}{2(ET)^2}. \] (4.10)
provided these second moments are finite. The first order approximation \( F^{(1)}_{ST} \) takes in the limit as \( \theta \downarrow 0 \) the same value as both \( F_{ST} \) and Slatkin's approximation \( F^{(0)}_{ST} \), but in contrast to \( F^{(0)}_{ST} \), which does not depend on \( \theta \), \( F^{(1)}_{ST} \) is (as a function of \( \theta \)) linear with derivative equal to the derivative of \( F_{ST} \) at \( \theta = 0 \). Retaining more terms in the Taylor expansion of \( F_{ST} \) would yield other approximations for \( F_{ST} \).

In Chapter 2 (Theorems 2.4, 2.5 and 2.6) we have derived systems of linear equations for the moment-generating function, the mean and the second moment of the coalescence time of a sample of genes from a subdivided population. In Chapter 3 we have solved these equations for a sample of two genes under a range of models of population structure. These results now enable us to calculate the values of \( F_{ST} \) and its approximations \( F^{(0)}_{ST} \) and \( F^{(1)}_{ST} \) explicitly under the various models of population structure considered in Chapter 3, and to examine the dependence of \( F_{ST} \) on the different parameters of population structure and on the mutation rate. In Section 2 we do this for the symmetric island and stepping-stone models, in Section 3 for the "less symmetric" structures introduced in Section 3.3.

### 4.2 The symmetric case

In this section, the analytical values of \( F_{ST} \) and its approximations are calculated under the symmetric island model (Subsection 1) and under the finite and infinite stepping-stone models in one and two dimensions (Subsections 2 to 5). For each model, figures illustrate how \( F_{ST} \) depends on the various parameters of the model and the accuracy of both Slatkin's and the first order approximation for \( F_{ST} \) is examined. Particular attention is paid to the dependence of \( F_{ST} \) on the neutral mutation rate, which is studied separately in Subsection 6. We show that the strength of the latter dependence varies a lot between the different models of population structure, which also explains the variation among population structures in the accuracy of the approximations for \( F_{ST} \). In the discussion (Subsection 7) some of the observations made for the various models are compared and some general properties of \( F_{ST} \) and its approximations are highlighted. Some criteria are proposed for Slatkin's approximation to be an accurate approximation for \( F_{ST} \).

In the symmetric island and stepping-stone models, all subpopulations are identical with respect to size and migration pattern, and all non-zero migration rates are equal. Recall that according to the notation used in Chapter 3, \( M \) is (for all models studied in this section) the total scaled migration rate out of each subpopulation, while \( \theta \) is the scaled mutation rate. Time-scaling is such that two genes in a single colony have coalescence rate 1 (working backwards in time). Each gene leaves its subpopulation at rate \( M/2 \) and mutates at rate \( \theta/2 \). If the subpopulations each contain \( 2N \) genes and evolve according to the neutral Wright-Fisher model, time-scaling is in units of \( 2N \) generations, \( M = \lim_{N \to \infty} (4Nm) \) and \( \theta = \lim_{N \to \infty} (4N\mu) \), where \( m \) is the proportion
Chapter 4: SUBPOPULATION DIFFERENTIATION

of each subpopulation that is exchanged for immigrants each generation (the specific migration pattern determines which subpopulations the immigrant genes come from) and where $\mu$ is the probability of mutation per gene per generation. So under the Wright-Fisher model, $M$ and $\theta$ are twice the number of migrant genes and twice the expected number of mutant genes, respectively, per subpopulation\(^5\) per generation. Under other models for reproduction, the time-scaling and hence the relationship between $M$ and $m$ and between $\theta$ and $\mu$ may be different.

For many loci it appears that mutation parameters $\theta$ in the range from 0 to about 1 are believed to be plausible (where a $\theta$ value of order 1 is in fact considered as very high). For this reason we focus on this range in most figures.

Note that an $F_{ST}$ value is specific to the collection of subpopulations considered, which does not necessarily include all subpopulations. There are two major types of $F_{ST}$ value that appear in the literature, and the distinction is not always clearly made. For island models, an $F_{ST}$ value is usually a global value, in which $\bar{\theta}$, $ET$ and $E[T^2]$ assume sampling of pairs of genes from the total population. The collection of subpopulations considered is in this case the collection of all subpopulations. For stepping-stone models, $F_{ST}$ values are more often pairwise values, in which $\bar{\theta}$, $ET$ and $E[T^2]$ are averages over a pair of subpopulations a specified distance apart. See also the footnote on page 73. Because in pairwise $F_{ST}$ values the contribution of genes from a single subpopulation (which tend to be more similar) has more weight than in global ones, pairwise $F_{ST}$ values are systematically lower than the corresponding global $F_{ST}$ values.

4.2.1 The symmetric island model

Denoting by $f_0$ the probability of identity by descent of two genes from a single subpopulation and by $f_1$ that of two genes from different subpopulations, their values are the moment-generating functions of the coalescence times $T_0$ and $T_1$ of, respectively, two genes from a single subpopulation and two genes from different subpopulations, evaluated in $s = \theta$ (cf. equations (4.4) and (4.5)). These moment-generating functions were calculated in Chapter 3 and are given by equations (3.1) and (3.2). So,

\begin{align*}
f_0 &= E[e^{-\theta T_0}] = \frac{M + (n-1)\theta}{M + (nM + n - 1)\theta + (n-1)\theta^2} \quad (4.11) \\
f_1 &= E[e^{-\theta T_1}] = \frac{M}{M + (nM + n - 1)\theta + (n-1)\theta^2}. \quad (4.12)
\end{align*}

Two genes chosen uniformly at random from the total population are from the same colony with probability $1/n$ and from different colonies with probability $1 - 1/n$. Hence the probability that

\(^5\)For a panmictic population under the Wright-Fisher model, time is generally scaled in terms of the total population size, and the scaled mutation rate, also denoted by $\theta$, is twice the expected number of mutations in the total population per generation. For a subdivided population, it is standard (see, for example, Takahata 1988, Hudson 1990, Notohara 1990) to scale time by subpopulation rather than population size, with $\theta$ and $M$ defined as in this thesis.
two genes randomly sampled from the total population are identical by descent is

\[ f = \frac{1}{n} f_0 + \left( 1 - \frac{1}{n} \right) f_1 \]

\[ = \frac{nM + (n-1) \theta}{n (M + (nM + n-1) \theta + (n-1) \theta^2)^{\frac{1}{2}}} \] (4.14)

Substituting (4.11) and (4.14) into the definition of \( F_{ST} \), equation (4.3), we find the (global) value of \( F_{ST} \):

\[ F_{ST} = \frac{1}{1 + M n^2/(n - 1)^2 + \theta n/(n - 1)} \] (4.15)

which was obtained earlier by Nei (1975) and Takahata (1983) by means of classical techniques.

The value of Slatkin's approximation \( F^{(0)}_{ST} \) is found from equation (4.8), where the mean coalescence time of two genes randomly sampled from the total population, \( E_T \), is calculated analogously to equation (4.13) from results (3.3) and (3.4). Alternatively, the value of \( F^{(0)}_{ST} \) is obtained by directly taking the limit of result (4.15) as \( \theta \rightarrow 0 \) (see the definition of \( F^{(0)}_{ST} \), equation (4.7)). We find:

\[ F^{(0)}_{ST} = \frac{1}{1 + M n^2/(n - 1)^2} \] (4.16)

(Slatkin 1991 and references therein). The first order approximation \( F^{(1)}_{ST} \) is calculated similarly, either from equation (4.10) and results (3.5) and (3.6), or directly from the definition of \( F^{(1)}_{ST} \), equation (4.9), and the result for \( F_{ST} \), equation (4.15). We obtain:

\[ F^{(1)}_{ST} = \frac{1}{1 + M n^2/(n - 1)^2} - \theta \frac{n/(n - 1)}{\{1 + M n^2/(n - 1)^2\}^2} \] (4.17)

We note that, as the number of colonies becomes large, Slatkin's approximation \( F^{(0)}_{ST} \) approaches \( 1/(1 + M) \), which is the well-known value first given by Wright (1931). This value is understood intuitively by noting that when there are very many colonies, two genes from different colonies need such a long time to get into a single colony, that (even for small mutation rates) a mutation will (almost) certainly have occurred before the two genes even get the possibility to coalesce. This means that the probability of identity by descent of two genes from different subpopulations is approximately zero. As in addition two genes randomly sampled from the total population are (almost) certainly taken from different colonies, the value of \( F_{ST} \) reduces to \( f_0 \), the probability that two genes from the same subpopulation are identical by descent. For the reasons just described and small mutation rates, this is (approximately) the probability that the two genes coalesce before a migration takes place. Knowing that the two genes coalesce at rate 1 and migrate at rate \( M/2 \) each, this probability is \( 1/(1 + M) \).

Figure 4.1 shows \( F_{ST} \) (given by equation (4.15)) as a function of the scaled migration rate \( M \), for different values of the scaled mutation rate \( \theta \), for an island model consisting of two subpopulations (figure 4.1a) and 20 subpopulations (figure 4.1b), respectively. For more than 20 subpopulations, the picture is almost identical to that in figure 4.1b. The thick line in both figures is Slatkin's
Chapter 4: SUBPOPULATION DIFFERENTIATION

Figure 4.1: $F_{ST}$ values as a function of the scaled migration rate, under the symmetric island model of population structure.

The thick solid line is the value of Slatkin's approximation $F^{(0)}_{ST}$, the thin lines are the values of $F_{ST}$ for, respectively, a scaled mutation rate $\theta = 0.1$ (---), $\theta = 0.5$ (---) and $\theta = 1$ (----). In figure a, there are $n = 2$ subpopulations, in figure b, the population is divided into $n = 20$ subpopulations.

approximation $F^{(0)}_{ST}$, which is the limit of $F_{ST}$ as $\theta \downarrow 0$. As $M$ tends to zero, $F_{ST}$ approaches its value under complete isolation of the subpopulations, namely $1/(1 + \frac{\theta}{n-\theta})$. Indeed, when the subpopulations are completely isolated, two genes from different subpopulations cannot be identical by descent ($f_1 = 0$) as they do not have a common ancestor, while two genes from the same subpopulation have probability $f_0 = 1/(1 + \theta)$ of being identical by descent (as this is the probability that their lineages coalesce before a mutation occurs on either of them). As is expected intuitively, the amount of genetic differentiation among colonies (as measured by $F_{ST}$) rapidly decreases as there is more gene flow between them (i.e. gene flow exercises a homogenizing effect on the genetic composition of the population, while isolation promotes differentiation among subpopulations). Analytically, this fast decrease of $F_{ST}$ is due to the rapid decrease of $f_0 - \bar{f}$ with increasing migration rate: as $M$ increases, it rapidly makes less difference (as far as probabilities of identity by descent are concerned) whether a pair of genes is sampled from a single colony or from the total population ($f_0$ is always larger than $\bar{f}$; with increasing $M$, $f_0$ decreases and $\bar{f}$ increases). As is seen from comparison of figure 4.1a with figure 4.1b, the value of $F_{ST}$ drops more rapidly for a small number of colonies.

Mutation rates in natural populations are usually very small, while migration rates are, in many species, much higher (Slatkin 1985). It is seen from figure 4.1 that in all those cases Slatkin's approximation $F^{(0)}_{ST}$ will be very good (if the symmetric island model is an appropriate description of the structure of the population), becoming increasingly accurate as the migration rate increases relative to the mutation rate. Only for large values of $\theta$ ($\theta$ larger than 0.1, which requires a very high value of the mutation rate $\mu$ or an unusually large effective subpopulation size), combined
with little gene flow, will there be a substantial difference between Slatkin's approximation and the exact $F_{ST}$ value.

Figure 4.2 shows the performance of the first order approximation $F^{(1)}_{ST}$ with increasing scaled migration rate, for two different values of the scaled mutation rate and for two different numbers of islands. For comparison, Slatkin's approximation $F^{(0)}_{ST}$ is included again in these plots (dotted line). Except for very small migration rates (roughly speaking, $m < \mu$, or in terms of the scaled rates: $M < \theta$), $F^{(1)}_{ST}$ is remarkably close to the exact $F_{ST}$ value. Note in particular that for $\theta = 0.1$, the graphs of $F_{ST}$ and its first order approximation $F^{(1)}_{ST}$ coincide in both figures 4.2a and 4.2b. As was the case for $F^{(0)}_{ST}$, the accuracy of $F^{(1)}_{ST}$ increases with increasing migration rate (for fixed mutation rate). The accuracy of both approximations $F^{(0)}_{ST}$ and $F^{(1)}_{ST}$ will be further discussed in Subsection 6, where we study and compare the dependence of $F_{ST}$ on the scaled mutation rate $\theta$ under different migration patterns.

The dependence of $F_{ST}$ and its approximations on the number of subpopulations is illustrated explicitly in figure 4.3. It is seen that $F_{ST}$ increases as the population is divided into a larger number of colonies. This is because, as $n$ increases, the probability, $f_1$, of identity by descent of two genes from different subpopulations, decreases much more quickly than that of two genes from the same colony, $f_0$; in addition, two genes chosen at random from the total population belong with increasing probability to different colonies and the contribution of genes from distinct colonies (which are on average more different than genes from the same subpopulation) towards $F_{ST}$ (via $\bar{f}$) gains weight. For small numbers of colonies, this dependence of $F_{ST}$ on the number of colonies is considerable. With more than about 10 or 20 subpopulations, $F_{ST}$ is very close to its limit as $n \to \infty$, namely $1/(1 + M + \theta)$, which is in turn close to Wright's value $1/(1 + M)$ if $\theta$ is sufficiently small.

For the sake of completeness, we also calculate the exact and approximate pairwise $F_{ST}$ values. Whilst the population is made up of $n$ colonies, the collection of subpopulations considered in the definition of $\bar{f}$ (following the definition of $F_{ST}$, equation (4.3)) is in that case a pair of colonies. Two genes sampled uniformly at random from the union of these two colonies are from the same colony with probability $1/2$. Hence

$$\bar{f} = \frac{f_0 + f_1}{2},$$

where $f_0$ and $f_1$ are given by equations (4.11) and (4.12), respectively. Denoting the $F_{ST}$ value of a pair of subpopulations by $F_{ST}(1)$, equation (4.3) gives

$$F_{ST}(1) = \frac{1}{1 + 2Mn/(n-1) + 2\theta}.$$

Taking the limit of this result as $\theta \downarrow 0$, the pairwise value of Slatkin's approximation is

$$F^{(0)}_{ST}(1) = \frac{1}{1 + 2Mn/(n-1)}.$$
Figure 4.2: $F_{ST}$ and its approximations under the symmetric island model. The values of $F_{ST}$ (thick lines) and its first order approximation $F_{ST}^{(1)}$ (thin lines) are plotted for both $\theta = 0.1$ (solid lines) and $\theta = 0.5$ (long-dashed lines). The dotted line shows $F_{ST}^{(0)}$. The number of colonies is $n = 2$ in figure a and $n = 20$ in figure b. In both figures, the graph of $F_{ST}^{(1)}$ for $\theta = 0.1$ is indistinguishable from that of $F_{ST}$. 
Figure 4.3: $F_{ST}$ values as a function of the number of subpopulations in the symmetric island model.

Values are plotted for migration rates $M = 0.1$ (diamonds), $M = 1$ (circles) and $M = 10$ (squares). The values of $F_{ST}^{(0)}$ (open symbols), $F_{ST}$ (filled symbols) and $F_{ST}^{(1)}$ (dotted symbols) are shown. The mutation rate is $\theta = 0.1$.

Alternatively this value could have been found from equation (4.8), where $ET$ is calculated analogously to equation (4.18) from results (3.3) and (3.4). The pairwise value of the first order approximation, denoted by $F_{ST}^{(1)}(1)$, is found from result (4.19) and equation (4.9), or from equation (4.10), calculating $E[T^2]$ analogously to equation (4.18) from results (3.5) and (3.6):

$$F_{ST}^{(1)}(1) = \frac{1}{1 + 2Mn/(n-1)} - \frac{2\theta}{(1 + 2Mn/(n-1))^2}.$$

### 4.2.2 The circular stepping-stone model

In calculating $F_{ST}$ values for a stepping-stone model, sampling is usually restricted to pairs of subpopulations at distance $d$, i.e. the collection of subpopulations considered is a pair of colonies $d$ steps apart. The probability of identity by descent of a pair of genes sampled at random from the union of these two subpopulations is

$$f = \frac{f_0 + f_d}{2}, \quad (4.20)$$

where $f_d$ is the probability of identity by descent of two genes at distance $d$. Substituting this in the definition of $F_{ST}$, equation (4.3), the $F_{ST}$ value of a pair of colonies at distance $d$ is

$$F_{ST}(d) = \frac{f_0 - f_d}{2 - (f_0 + f_d)}. \quad (4.21)$$

The value of $f_d$ is result (3.22) for the moment-generating function of the coalescence time $T_d$ of two genes $d$ colonies apart, with $s$ replaced by $\theta$. Substituting this result into (4.21), we find for
Chapter 4: SUBPOPULATION DIFFERENTIATION

\[ d = 1, \ldots, [n/2]: \]

\[ F_{ST}(d) = \frac{a_+(\theta) \lambda_+([n/2]-d) \{\lambda_+ (\theta)^d - M^d\} + a_- (\theta) \lambda_-([n/2]-d) \{\lambda_- (\theta)^d - M^d\}}{a_+(\theta) \lambda_+([n/2]-d) \{b_+ (\theta) \lambda_+ (\theta)^d - M^d\} + a_- (\theta) \lambda_-([n/2]-d) \{b_- (\theta) \lambda_- (\theta)^d - M^d\}} \]

(4.22)

where \( a_+ (\cdot), a_- (\cdot), \lambda_+ (\cdot) \) and \( \lambda_- (\cdot) \) are defined by equations (3.23) and (3.21) and where in addition

\[ b_+ (\theta) = 1 + 2\sqrt{(2M + \theta)\theta} \]

\[ b_- (\theta) = 1 - 2\sqrt{(2M + \theta)\theta}. \]

Alternatively, substitution of (3.24) into (4.21) gives the following equivalent expression for the \( F_{ST} \) value of a pair of colonies at distance \( d \):

\[ F_{ST}(d) = \frac{1}{n} \sum_{k=0}^{n-1} \frac{1 - \cos \frac{2\pi kd}{n}}{\theta + M (1 - \cos \frac{2\pi k}{n})} \]

for \( d = 1, \ldots, [n/2] \).

Taking averages as in (4.20), we have for Slatkin's approximation \( F_{ST}^{(0)} \), calculated for a pair of colonies \( d \) steps apart:

\[ F_{ST}^{(0)} (d) = \frac{E T_d - E T_0}{E T_d + E T_0} \]

(Slatkin 1991), while the first order approximation \( F_{ST}^{(1)} \) for a pair of colonies at distance \( d \) is given by

\[ F_{ST}^{(1)} (d) = \frac{E T_d - E T_0}{E T_d + E T_0} - \theta \cdot \frac{E[T_2] - E[T_2^0]}{(E T_d + E T_0)^2}. \]

(4.24)

Substituting results (3.25) and (3.26) into equations (4.23) and (4.24), we obtain for \( d = 1, \ldots, [n/2] \):

\[ F_{ST}^{(0)} (d) = \frac{1}{1 + 2M n/[d(n - d)]} \]

(4.25)

(Slatkin 1991) and

\[ F_{ST}^{(1)} (d) = \frac{1}{1 + 2M n/[d(n - d)]} - \frac{\theta n}{3} \cdot \frac{1 + 2/[d(n - d)]}{\{1 + 2M n/[d(n - d)]\}^{3/2}}. \]

(4.26)

In figure 4.4 the values of \( F_{ST}^{(0)} (d) \) (thick solid line) and \( F_{ST} (d) \), given by equations (4.25) and (4.22), are plotted against the scaled migration rate \( M \), for various values of the scaled mutation rate \( \theta \), the number of colonies \( n \) and the distance \( d \) between the pair of colonies considered. The value of \( F_{ST} (d) \) in the limit of complete isolation of the subpopulations (\( M \uparrow 0 \)) is the same in all four plots in figure 4.4, namely \( 1/(1 + 2\theta) \), independent of \( n \) and \( d \). This limiting value is different from that under the island model only because the \( F_{ST} \) values for the stepping-stone models are calculated for a pair of subpopulations, while \( F_{ST} \) for the island model is a global value using information on all subpopulations. Comparing figure 4.4a with figure 4.4b, and figure 4.4c
Figure 4.4: $F_{ST}$ values as a function of the scaled migration rate, for a pair of colonies in a circular stepping-stone model. Two colonies at distance $d$ on a circle with $n$ colonies are considered, where $d = 2$ and $n = 4$ (figure a), $d = 2$ and $n = 20$ (figure b), $d = 10$ and $n = 20$ (figure c), and $d = 10$ and $n = 100$ (figure d). The thick solid line is Slatkin's approximation $F_{ST}^{(0)}(d)$. The thin lines are the values of $F_{ST}(d)$ at scaled mutation rates $\theta = 1$ (dot-dashed line), $\theta = 0.5$ (long-dashed), $\theta = 0.1$ (dashed) and $\theta = 0.01$ (dotted; only shown in figures c and d).
with figure 4.4d, one sees that for a fixed distance \( d \), \( F_{ST} \) values fall off more slowly with increasing migration rate when there are more colonies. This means that (at the same scaled migration rate) two subpopulations ten steps apart in a circle consisting of 100 colonies, for instance, are genetically more different (as measured by \( F_{ST} \)) than two colonies 10 steps apart in a circular stepping-stone model with only 20 colonies. From comparison of figure 4.4b with figure 4.4c it is seen that \( F_{ST}(d) \) drops much more slowly with increasing \( M \) as the distance \( d \) between the two colonies under consideration is larger, leaving much higher \( F_{ST}(d) \) values for pairs of colonies a further distance apart. Hence colonies at a large distance from each other are much more genetically differentiated than colonies that are near to each other, which is what one expects intuitively.

From figure 4.4 it is also seen that Slatkin's approximation \( F_{ST}^{(0)}(d) \) is very accurate for two colonies a small distance apart in a circular stepping-stone model consisting of only few colonies, but becomes considerably less accurate when there are many colonies or the distance between the two colonies considered is large. This has to do with the fact that \( F_{ST}(d) \) decreases more rapidly with increasing mutation rate \( \theta \) as the distance \( d \) or the number of colonies \( n \) increase. In the case of a pair of colonies at distance 10 on a circle consisting of 20 colonies (figure 4.4c) with mutation rate \( \theta = 0.1 \), Slatkin's approximation may overrate the real \( F_{ST}(d) \) value by about 40%. For two subpopulations at distance 10 on a circle with 100 subpopulations (figure 4.4d) and \( \theta = 0.1 \), the discrepancy between \( F_{ST}^{(0)}(d) \) and \( F_{ST}(d) \) can be over 60%, relative to \( F_{ST}(d) \).

Because for a single model, global \( F_{ST} \) values are systematically higher than pairwise \( F_{ST} \) values (as \( F_{ST} \) values for pairs of subpopulations include a higher contribution of \( f_0 \) towards \( f \)), the values in figure 4.4 do not allow for comparison with those in figure 4.1 (symmetric island model). Nevertheless it is seen from these figures that Slatkin's approximation is in general less accurate for a circular stepping-stone model than it is for an island model.

The increase of \( F_{ST}(d) \) and both its approximations with the distance \( d \) is shown explicitly in figure 4.5, from which it is seen that this increase is very important for small distances, but becomes weak as the distance gets large. This is due to the very rapid (almost exponential) decrease of \( f_d \) with increasing \( d \) (noted earlier by Malécot 1948, 1975, and by Maruyama 1970), causing \( f_d \) to be close to zero and \( F_{ST}(d) \) to be close to \( f_0/(2 - f_0) \) as soon as \( d \) gets larger. This approach is faster as the colonies are more isolated. The increase of (measures of) genetic differentiation with increasing geographic distance is an intuitively evident and long known fact referred to by Wright (1943) as "isolation by distance" (see also Slatkin 1993 and references therein). Figure 4.5 also shows that at fairly low levels of gene flow (and with a large number of colonies), the first order approximation \( F_{ST}^{(1)}(d) \) is more inaccurate than \( F_{ST}^{(0)}(d) \) even for \( \theta \) as small as 0.01. We will return to the reasons for the inaccuracy of both approximations under this particular model in Subsection 6.

In figure 4.6, \( F_{ST}(d) \) and both its approximations are plotted against the number \( (n) \) of sub-
Chapter 4: SUBPOPULATION DIFFERENTIATION

Figure 4.5: Pairwise $F_{ST}$ values under the circular stepping-stone model, as a function of the distance between the two colonies under consideration.

Values of $F_{ST}(d)$ (diamonds) and of both its approximations $F_{ST}^{(0)}(d)$ (circles) and $F_{ST}^{(1)}(d)$ (squares) are shown for migration rate $M = 1$ (filled symbols) and for migration rate $M = 10$ (no fill). We have assumed $n = 100$ subpopulations on the circle and a mutation rate $\theta = 10^{-2}$.

populations on the circle, for two colonies a fixed distance $d = 10$ apart. As was the case for the island model, both $F_{ST}(d)$ and $F_{ST}^{(0)}(d)$ increase with increasing $n$ when $n$ is small, but soon become approximately independent of $n$, approaching their limiting values. These limiting values are the values of $F_{ST}(d)$ and $F_{ST}^{(0)}(d)$ under the infinite linear stepping-stone model which will be discussed in the next subsection. The first order approximation $F_{ST}^{(1)}(d)$ is accurate for small numbers of subpopulations but useless for large numbers of subpopulations. Note that except for small $n$, the discrepancy between $F_{ST}(d)$ and $F_{ST}^{(1)}(d)$ increases approximately linearly with increasing $n$.

We also calculate the global values of $F_{ST}$ and its approximations for this model, in which $\bar{f}$, $ET$ and $E[T^2]$ involve sampling from the total population rather than from a pair of subpopulations $d$ steps apart. These global $F_{ST}$ values will be used in Subsection 6 to compare the circular stepping-stone model with the symmetric island model.

The probability that two genes sampled uniformly at random from the total population are identical by descent is

$$
\bar{f} = \begin{cases} 
\frac{1}{n} f_0 + \frac{2}{n} \sum_{d=1}^{n-1} f_d + \frac{1}{n} f_\frac{n}{2} & \text{if } n \text{ is even} \\
\frac{1}{n} f_0 + \frac{2}{n} \sum_{d=1}^{n-1} f_d & \text{if } n \text{ is odd}
\end{cases}
$$

(4.27)
Figure 4.6: Pairwise $F_{ST}$ values under the circular stepping-stone model, as a function of the number of subpopulations on the circle.

Values of $F_{ST}(d)$ and both its approximations are shown for two colonies $d = 10$ steps apart, for two values of the scaled migration rate: $M = 1$ (circles) and $M = 10$ (squares). The scaled mutation rate is $\theta = 10^{-2}$. The open, filled and dotted symbols represent respectively the values of $F^{(0)}_{ST}(d)$, $F_{ST}(d)$ and $F^{(1)}_{ST}(d)$.

which is calculated most easily from expression (3.24) for the moment-generating function of $T_d$:

$$f = \frac{1}{n^2} \sum_{k=0}^{n-1} \frac{\cos \frac{2\pi kd}{n}}{1 + \frac{n}{n} \sum_{k=0}^{n-1} \frac{1}{\theta + M \left(1 - \cos \frac{2\pi k}{n}\right)}}$$

recalling that

$$\sum_{d=0}^{n-1} \frac{2\pi kd}{n} = \begin{cases} 0 & \text{for } k = 1, \ldots, n-1 \\ n & \text{for } k = 0. \end{cases}$$

Result (4.28) is the limit of infinite subpopulation size of a result in Maruyama (1970). Substituting (3.24) with $d = 0$ and (4.28) into the definition of $F_{ST}$, equation (4.3), we obtain the global value of $F_{ST}$:

$$F_{ST} = \frac{\sum_{k=1}^{n-1} \frac{1}{\theta + M \left(1 - \cos \frac{2\pi k}{n}\right)}}{n + \sum_{k=1}^{n-1} \frac{1}{\theta + M \left(1 - \cos \frac{2\pi k}{n}\right)}}.$$
Chapter 4: SUBPOPULATION DIFFERENTIATION

We can evaluate the sums in this result by noting that equating expressions (3.22) and (3.24), with \( d = 0 \), gives the following identity:

\[
\frac{1}{n} \sum_{k=0}^{n-1} \frac{1}{\theta + M (1 - \cos \frac{2\pi k}{n})} = \frac{1}{\sqrt{(2M + \theta)\theta}} \cdot \frac{a_+(\theta)\lambda_+(\theta)^{n/2} + a_-(\theta)\lambda_-(\theta)^{n/2}}{a_+(\theta)\lambda_+(\theta)^{n/2} - a_-(\theta)\lambda_-(\theta)^{n/2}}
\]

where \( a_+(\cdot) \), \( a_-(\cdot) \), \( \lambda_+(\cdot) \) and \( \lambda_-(\cdot) \) are given by equations (3.23) and (3.21). Substituting this identity into the global value of \( F_{ST} \), result (4.29), we find the following explicit expression:

\[
F_{ST} = \frac{(n\theta - \sqrt{(2M + \theta)\theta}) a_+(\theta)\lambda_+(\theta)^{n/2} + (n\theta + \sqrt{(2M + \theta)\theta}) a_-(\theta)\lambda_-(\theta)^{n/2}}{(n\theta - (1 - n\theta)\sqrt{(2M + \theta)\theta}) a_+(\theta)\lambda_+(\theta)^{n/2} + (n\theta + (1 - n\theta)\sqrt{(2M + \theta)\theta}) a_-(\theta)\lambda_-(\theta)^{n/2}}
\]

(4.30)

The global values of Slatkin's approximation and of the first order approximation for \( F_{ST} \) are found from equations (4.8) and (4.10), where \( ET \) and \( E[T^2] \) are calculated analogously to equation (4.27). Alternatively they can be obtained from result (4.30) by directly evaluating equations (4.7) and (4.9). The results are:

\[
F_{ST}^{(0)} = \frac{n^2 - 1}{n^2 - 1 + 6nM}
\]

(4.31)

\[
F_{ST}^{(1)} = \frac{n^2 - 1}{n^2 - 1 + 6nM} \cdot \frac{\theta n^4 + 10n^2 - 11}{5(n^2 - 1 + 6nM)^2}
\]

(4.32)

4.2.3 The infinite linear stepping-stone model

The \( F_{ST} \) value of a pair of colonies at distance \( d \), denoted by \( F_{ST}(d) \), is calculated by substituting result (3.27) for the moment-generating function of the coalescence time \( T_d \) of a pair of genes at distance \( d \), evaluated in \( s = \theta \), into equation (4.21):

\[
F_{ST}(d) = \frac{M^d - \left(M + \theta - \sqrt{(2M + \theta)\theta}\right)^d}{M^d - \left(M + \theta - \sqrt{(2M + \theta)\theta}\right)^d + 2M^d \cdot \sqrt{(2M + \theta)\theta}}
\]

(4.33)

for \( d = 1, 2, \ldots \). An equivalent expression in terms of an integral can be found from result (3.28).

As the mean coalescence time of any two individuals under this model is infinite (result (3.29)), Slatkin's recipe for \( F_{ST}^{(0)}(d) \), equation (4.23), would give an undefined expression. However, the value of \( F_{ST}^{(0)}(d) \) can be found by directly calculating the limit of \( F_{ST}(d) \), result (4.33), as \( \theta \downarrow 0 \) (see the definition of \( F_{ST}^{(0)} \), equation (4.7)). Using l'Hôpital's rule, we obtain:

\[
F_{ST}^{(0)}(d) = \frac{1}{1 + 2M/d}
\]

(4.34)

for \( d = 1, 2, \ldots \). This value was found by Slatkin (1991) as an approximation valid for a circular stepping-stone model with a large number of colonies.
Figure 4.7: Pairwise $F_{ST}$ values as a function of the scaled migration rate, under the infinite linear stepping-stone model.

Values of $F_{ST}(d)$ (thin lines) and its approximation $F_{ST}^{(0)}(d)$ (thick line) are shown for two colonies at distance $d = 10$ (figure a) and distance $d = 100$ (figure b). The $F_{ST}(d)$ values were calculated for scaled mutation rates $\theta = 1$ (dot-dashed line), $\theta = 0.1$ (dashed), $\theta = 0.01$ (dotted), $\theta = 10^{-3}$ (solid) and $\theta = 10^{-4}$ (long-dashed; figure b only). The first order approximation $F_{ST}^{(1)}(d)$ cannot be considered for this model, as its value is $-\infty$.

To find the value of the first order approximation $F_{ST}^{(1)}(d)$, defined by equation (4.9), for a pair of colonies at distance $d$, we similarly calculated the limit of the derivative of $F_{ST}(d)$ with respect to $\theta$, as $\theta \downarrow 0$, directly from result (4.33) and found (using l'Hopital's rule) that

$$\lim_{\theta \downarrow 0} \frac{\partial}{\partial \theta} F_{ST}(d) = -\infty$$

for $d = 1, 2, \ldots$. As a consequence,

$$F_{ST}^{(1)}(d) = -\infty$$

and is of no use as an approximation for $F_{ST}(d)$. Further implications of (4.35) towards $F_{ST}(d)$ and towards the accuracy of $F_{ST}^{(0)}(d)$ will be discussed in Subsection 6.

Figure 4.7 shows $F_{ST}(d)$ and $F_{ST}^{(0)}(d)$ (thick solid line) plotted against the migration rate, for different values of the mutation rate. In figure 4.7a, a pair of colonies at distance $d = 10$ is considered. In figure 4.7b, the distance between the two colonies is $d = 100$. As was the case for the circular stepping-stone model, $F_{ST}(d)$ starts from its limiting value $1/(1 + 2\theta)$, independent of $d$, as $M \downarrow 0$. Its rate of decrease with increasing $M$ is slower as the distance $d$ between the two colonies considered is larger. With increasing $M$, both the absolute discrepancy $F_{ST}^{(0)}(d) - F_{ST}(d)$ and the relative discrepancy $(F_{ST}^{(0)}(d) - F_{ST}(d))/F_{ST}(d)$ between $F_{ST}(d)$ and its approximation $F_{ST}^{(0)}(d)$ initially increase to a maximum, after which they decrease. For the relative discrepancy, this decrease tends to be very slow. As the distance $d$ is larger, the maximum discrepancies are larger and are reached at larger values of the migration rate $M$. For two colonies at distance 100 and for $\theta = 0.01$, the maximum relative discrepancy between the real $F_{ST}(d)$ value and its
Figure 4.8: Pairwise $F_{ST}$ values under the infinite linear stepping-stone model, as a function of the distance between the pair of colonies considered. Values of $F_{ST}(d)$ (diamonds) and its approximation $F_{ST}^{(0)}(d)$ (circles) are plotted for migration rates $M = 1$ (filled symbols) and $M = 25$ (no fill). A mutation rate of $\theta = 10^{-3}$ is assumed. One symbol is plotted every 10th value of $d$.

approximation is as much as 67% and is reached about $M = 31$ (at that migration rate the values of $F_{ST}^{(0)}(d)$ and $F_{ST}(d)$ are 0.617 and 0.369, respectively). For $\theta = 0.001$, the maximum relative discrepancy is 18%, reached about $M = 41$. The $M$-value at which the maximum discrepancy is reached increases with decreasing mutation rate $\theta$.

Figure 4.8 shows the rapid increase of the amount of subpopulation differentiation (as measured by $F_{ST}(d)$) with increasing distance $d$ between the two colonies considered. This rapid increase is caused by the exponential decrease of the probability of identity by descent of two genes at distance $d$, first noted by Malécot (1948). It is seen that this increase is more rapid as there is less gene flow between colonies. As the distance $d$ becomes large, $F_{ST}(d)$ approaches its limiting value $1/(1 + 2\sqrt{(2M + \theta)\theta})$. Its approximation $F_{ST}^{(0)}(d)$ always converges to 1, as $d$ grows to infinity. Hence for a fixed scaled mutation rate $\theta$, the discrepancy between $F_{ST}(d)$ and $F_{ST}^{(0)}(d)$ keeps increasing with increasing distance, with limiting value $2\sqrt{(2M + \theta)\theta}/(1 + 2\sqrt{(2M + \theta)\theta})$ as $d \to \infty$. This limiting discrepancy can be very large if the scaled migration rate $M$ is large. In that case, however, the increase of the discrepancy between $F_{ST}(d)$ and $F_{ST}^{(0)}(d)$ to its large limiting value is slow and the limiting value will only be approached at very large distances $d$. 
4.2.4 Two-dimensional stepping-stone model on a torus

Values of $F_{ST}$ for this model are usually pairwise values and are a function of the numbers of steps, $d_1$ and $d_2$, separating the two colonies in the respective dimensions of the torus. The exact $F_{ST}$ value, $F_{ST}(d_1, d_2)$, of a pair of colonies at “distance” $(d_1, d_2)$ on a $K \times L$ torus is calculated from result (3.40) according to equation (4.21) with $0$ replaced by $(0,0)$ and $d$ by $(d_1, d_2)$:

$$F_{ST}(d_1, d_2) = \frac{\sum_{v=0}^{K-1} \sum_{w=0}^{L-1} \frac{1 - \cos(2\pi d_1/K)\cos(2\pi d_2/L)}{M + 4M + M\cos(2\pi d_1/K) + \cos(2\pi d_2/L)/2}}{2KL + \sum_{v=0}^{K-1} \sum_{w=0}^{L-1} \frac{1 - \cos(2\pi d_1/K)\cos(2\pi d_2/L)}{M + 4M + M\cos(2\pi d_1/K) + \cos(2\pi d_2/L)/2}}$$

(4.36)

for $d_1 = 0,\ldots,[K/2]$ and $d_2 = 0,\ldots,[L/2]$ with $(d_1, d_2) \neq (0,0)$.

The values of Slatkin’s approximation $F_{ST}^{(0)}(d_1, d_2)$ and the first order approximation $F_{ST}^{(1)}(d_1, d_2)$ are found analogously according to equations (4.23) and (4.24), using results (3.41) and (3.42), or are found by letting $\theta = 0$ in result (4.36) for $F_{ST}(d_1, d_2)$ and in its derivative with respect to $\theta$.

We obtain for $d_1 = 0,\ldots,[K/2]$ and $d_2 = 0,\ldots,[L/2]$ with $(d_1, d_2) \neq (0,0)$:

$$F_{ST}^{(0)}(d_1, d_2) = \frac{\sum_{(v,w)\in S_0} 1 - \frac{\cos(2\pi d_1/K)\cos(2\pi d_2/L)}{M + \cos(2\pi d_1/K) + \cos(2\pi d_2/L)/2}}{2KL + \sum_{(v,w)\in S_0} 1 - \frac{\cos(2\pi d_1/K)\cos(2\pi d_2/L)}{M + \cos(2\pi d_1/K) + \cos(2\pi d_2/L)/2}}$$

(Slatkin 1991, 1993) and

$$F_{ST}^{(1)}(d_1, d_2) = F_{ST}^{(0)}(d_1, d_2) - \theta \cdot \frac{2KL \sum_{(v,w)\in S_0} 1 - \frac{\cos(2\pi d_1/K)\cos(2\pi d_2/L)}{M + \cos(2\pi d_1/K) + \cos(2\pi d_2/L)/2}}{\left(2KL + \sum_{(v,w)\in S_0} 1 - \frac{\cos(2\pi d_1/K)\cos(2\pi d_2/L)}{M + \cos(2\pi d_1/K) + \cos(2\pi d_2/L)/2}\right)^2}$$

where, as in Subsection 3.2.3, $S_0 = \{0,\ldots,K-1\} \times \{0,\ldots,L-1\}\setminus\{(0,0)\}$.

An interesting feature of this finite two-dimensional stepping-stone model is that it allows us to investigate the dependence of $F_{ST}$ on habitat shape. Figure 4.9 shows the values of $F_{ST}(d_1, d_2)$ and both its approximations for a torus with 900 colonies in total, whose shape varies from a $1 \times 900$ one-dimensional array on a circle to a $30 \times 30$ “square” torus. The $F_{ST}$ values shown are those of a pair of colonies 15 steps apart in the dimension with the larger number of colonies, for three different values of the migration rate. The mutation rate is $\theta = 10^{-3}$. It is seen that the amount of genetic differentiation between the two colonies, as measured by $F_{ST}(d_1, d_2)$, is much larger in an essentially one-dimensional (i.e. one-dimensional or very narrow two-dimensional) habitat and decreases very rapidly as the width of the habitat increases. Once the habitat is truly two-dimensional (which appears to be from about the $10 \times 90$ torus onwards, say), $F_{ST}(d_1, d_2)$ depends only weakly on the precise shape of the torus and approaches the value for the square torus. Similar observations were made by Crow and Aoki (1984) for the global $F_{ST}$ values (calculated numerically rather than analytically). The large $F_{ST}$ values in the long and narrow habitats towards the left-hand side of the $X$-axis in figure 4.9 are really caused by the narrow width of the habitat, rather than by its large length: two colonies 15 steps apart on a circle consisting of only 30 colonies with scaled migration rate $M$ going out of each colony (data not shown) still
Figure 4.9: Pairwise $F_{ST}$ values for a stepping-stone model on a torus, for various shapes of the habitat. Values of $F_{ST}(d_1, d_2)$ and both its approximations are shown for a pair of colonies at distance $(d_1, d_2) = (0, 15)$ on a torus with $K \times L = 900$ colonies in total. The shape of the torus ranges from a circle ($K = 1, L = 900$) to a "square torus" ($K = L = 30$). Values are plotted for three different migration rates: $M = 1$ (diamonds), $M = 5$ (circles) and $M = 10$ (squares). The symbols, joined by a solid line, represent the value of $F_{ST}(d_1, d_2)$. Slatkin's approximation $F_{ST}^{(0)}(d_1, d_2)$ is indicated by a dotted line, the linear approximation $F_{ST}^{(1)}(d_1, d_2)$ by a long-dashed line. The mutation rate is $\theta = 10^{-3}$. For all true toruses ($K \geq 2$), the scaled migration rate is $M$, of which rate $M/2$ is in each of the two dimensions. For the $1 \times 900$ circle, we have plotted the $F_{ST}$ values for scaled migration rate $M/2$ in its only one dimension, to maintain the continuity of the graph. Using a scaled migration rate of $M$ for the circle would lead to lower $F_{ST}$ values than shown here.
Chapter 4: SUBPOPULATION DIFFERENTIATION

Figure 4.10: The decrease of $F_{ST}(d_1, d_2)$ with increasing scaled migration rate in a stepping-stone model on a torus, compared between various shapes of the habitat.

We have plotted the exact $F_{ST}(d_1, d_2)$ values of a pair of colonies $(d_1, d_2) = (0, 15)$ steps apart in a torus with 900 subpopulations in total. The mutation rate is $\theta = 10^{-3}$. The solid line represents a square torus $(30 \times 30)$, the dotted line a $10 \times 90$ torus, the long-dashed line a $3 \times 300$ torus and the dot-dashed line a $1 \times 900$ circle. In all cases (including the circle), $M$ is the total (scaled) migration rate out of each colony.

We have a substantially higher $F_{ST}$ value than two colonies at distance $(0, 15)$ on a $30 \times 30$ torus with scaled migration rate $M$, although in the latter situation the number of subpopulations is 30 times larger. We conclude that with regard to $F_{ST}$ as a measure of subpopulation differentiation, two-dimensional dispersal of alleles has a much stronger homogenizing effect on the population than has one-dimensional dispersal. It is also observed that both approximations $F_{ST}^{(0)}(d_1, d_2)$ and $F_{ST}^{(1)}(d_1, d_2)$ are very accurate for habitats that are truly two-dimensional but less accurate for essentially one-dimensional habitats. For a truly two-dimensional habitat, the first order approximation is for small scaled mutation rates more accurate than Slatkin's approximation (unless the number of colonies is very large, as we will see later). When the habitat is narrow, the first order approximation becomes unreliable.

We also see from figure 4.9 that the dependence of $F_{ST}$ on habitat shape is more important for somewhat larger values of $M$. For small migration rates, subpopulations will be very different under any model of population structure (when the population is subject to neutral mutation). In a two-dimensional habitat, an increasing level of gene flow will soon decrease differences among subpopulations, while in narrow habitats $F_{ST}$ decreases only slowly with increasing $M$ and a substantial amount of subpopulation differentiation is preserved even at relatively large scaled migration rates. This is illustrated explicitly in figure 4.10. At very large migration rates, there
Figure 4.11: Pairwise $F_{ST}$ values under a stepping-stone model on a torus, as a function of the distance between the pair of colonies considered. The values of $F_{ST}(d_1, d_2)$ (filled symbols), Slatkin's approximation $F_{ST}^{(0)}(d_1, d_2)$ (no fill) and the first order approximation $F_{ST}^{(1)}(d_1, d_2)$ (dotted symbols) of a pair of colonies in a 100 x 100 torus are shown. In figure a, $d$ is the number of steps separating the two colonies in one dimension ($d_1 = d$), while the distance in the other dimension is zero ($d_2 = 0$). In figure b, the two colonies are $d$ steps apart in each dimension, i.e. $d_1 = d_2 = d$. Migration rates of $M = 1$ (diamonds), $M = 5$ (circles) and $M = 20$ (squares) are considered. The mutation rate is $\theta = 10^{-3}$.

will of course be little differentiation even under a one-dimensional stepping-stone model.

Figure 4.11 shows the dependence of the $F_{ST}$ values on the distance between the two colonies considered in a 100 x 100 torus. In figure 4.11a, the distance between the two subpopulations varies in only one dimension (assuming distance zero in the other) and the values of $F_{ST}(d, 0)$ and its approximations are plotted against $d$. In figure 4.11b, the two colonies considered are $d$ steps apart in each dimension, i.e. their distance is $(d, d)$. It is seen that, except for small distances, $F_{ST}$ depends only weakly on the distance and two colonies separated by $d$ steps in each dimension are hardly any more differentiated than two colonies $d$ steps apart only in one dimension. For small values of $d$ the increase of $F_{ST}$ with increasing distance is important, but much less substantial than in a one-dimensional stepping-stone model (as can be seen by comparing figure 4.11 with figure 4.5). Both Slatkin's approximation and the first order approximation become less accurate as $d$ increases.

Figure 4.12 shows that for a pair of colonies a fixed distance apart on a square torus, the value of $F_{ST}(d_1, d_2)$ is nearly independent of the number of colonies on the torus (unless both the distance and the number of subpopulations are very small). Both Slatkin’s approximation and the first order approximation become less accurate as the number of subpopulations increases. This effect is more important for the first order approximation. However, comparison with figure 4.6 shows that the decrease of accuracy of the first order approximation is much less rapid than under the one-dimensional stepping-stone model.
Figure 4.12: Pairwise $F_{ST}$ values under a stepping-stone model on a $K \times K$ torus, as a function of $K$.

The values shown are those of $F_{ST}(5, 5)$ (filled symbols), Slatkin's approximation $F_{ST}^{(0)}(5, 5)$ (no fill) and the first order approximation $F_{ST}^{(1)}(5, 5)$ (dotted symbols) of two colonies that are 5 steps apart in each dimension of a $K \times K$ torus. Three values of the scaled migration rate are shown: $M = 1$ (diamonds), $M = 5$ (circles) and $M = 10$ (squares). All values were calculated for a scaled mutation rate $\theta = 10^{-2}$.

In Subsection 6 we will also use the global values of $F_{ST}$ and its approximations under this model. The probability of identity by descent of two genes drawn at random from the total population is, with result (3.40),

$$f = \frac{1}{KL} \sum_{v=0}^{K-1} \sum_{w=0}^{L-1} \frac{\sum_{d_1=0}^{K-1} \cos(2d_1v/K) \sum_{d_2=0}^{L-1} \cos(2d_2w/L)}{M + \theta - M[\cos(2v/K) + \cos(2w/L)]/2}$$

$$= \frac{1}{\theta \{KL + \sum_{v=0}^{K-1} \sum_{w=0}^{L-1} M + \theta - M[\cos(2v/K) + \cos(2w/L)]/2 \}}$$

for $K$ and $L$ even or odd. Substituting this and result (3.40) with $d_1 = d_2 = 0$ into equation (4.3), we find the global value of $F_{ST}$:

$$F_{ST} = \frac{\sum_{(v,w) \in S_0} \frac{1}{M + \theta - M[\cos(2v/K) + \cos(2w/L)]/2}}{KL + \sum_{(v,w) \in S_0} \frac{1}{M + \theta - M[\cos(2v/K) + \cos(2w/L)]/2}}.$$

(4.37)

Taking the limit as $\theta \downarrow 0$ of this result and of its derivative with respect to $\theta$, we find the global values of the approximations $F_{ST}^{(0)}$ and $F_{ST}^{(1)}$ for $F_{ST}$ from equations (4.7) and (4.9), respectively:

$$F_{ST}^{(0)} = \frac{\sum_{(v,w) \in S_0} \frac{1}{1 - [\cos(2v/K) + \cos(2w/L)]/2}}{KL + \sum_{(v,w) \in S_0} \frac{1}{1 - [\cos(2v/K) + \cos(2w/L)]/2}}.$$

(4.38)

$$F_{ST}^{(1)} = F_{ST}^{(0)} - \theta \frac{KL \sum_{(v,w) \in S_0} \left\{1 - [\cos(2v/K) + \cos(2w/L)]/2 \right\}}{KL + \sum_{(v,w) \in S_0} \frac{1}{1 - [\cos(2v/K) + \cos(2w/L)]/2}}.$$

(4.39)
The exact $F_{ST}$ value of a pair of subpopulations $d_1$ and $d_2$ steps apart in the respective dimensions of the lattice, denoted by $F_{ST}(d_1, d_2)$, is calculated according to equation (4.21) with $d$ replaced by $(d_1, d_2)$ and 0 by $(0, 0)$, from result (3.49) in terms of double integrals or from result (3.51) in terms of single integrals:

$$ F_{ST}(d_1, d_2) = \frac{\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1 - \cos(d_1 x) \cos(d_2 y)}{M + \theta - M(\cos x + \cos y) / 2} \, dx \, dy}{8\pi^2 M + \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1 - \cos(d_1 x) \cos(d_2 y)}{M + \theta - M(\cos x + \cos y) / 2} \, dx \, dy} \tag{4.40} $$

$$ = \frac{\int_{0}^{2\pi} \frac{1 - \cos(d_1 x) (2 + 2\theta/M - \cos x - \sqrt{(2 + 2\theta/M - \cos x)^2 - 1})^2}{\sqrt{(2 + 2\theta/M - \cos x)^2 - 1}} \, dx}{\pi M + \int_{0}^{2\pi} \frac{1 - \cos(d_1 x) (2 + 2\theta/M - \cos x - \sqrt{(2 + 2\theta/M - \cos x)^2 - 1})^2}{\sqrt{(2 + 2\theta/M - \cos x)^2 - 1}} \, dx} \tag{4.41} $$

for $d_1, d_2 \in \mathbb{N}$ with $(d_1, d_2) \neq (0, 0)$.

Because the mean and the second moment of the coalescence time of any pair of genes are infinite (results (3.52) and (3.53)), the values of the approximations $F_{ST}^{(0)}(d_1, d_2)$ and $F_{ST}^{(1)}(d_1, d_2)$ cannot be obtained from equations (4.23) and (4.24), but have to be calculated directly from $F_{ST}(d_1, d_2)$ according to their definitions, equations (4.7) and (4.9). The integrand of the double integral in result (4.40) is non-negative and increases as $\theta$ decreases. The same is true for the integrand of the single integral in result (4.41), as

$$ 1 - \cos(d_1 x) \left(2 + 2\theta/M - \cos x - \sqrt{(2 + 2\theta/M - \cos x)^2 - 1}\right)^2 \sqrt{(2 + 2\theta/M - \cos x)^2 - 1} \, dx \, dy $$

(see (3.50), with $M$ replaced by $M / 2$ and with $s = \theta + M / 2(1 - \cos x)$), for $M > 0, \theta > 0$ and $x \in \mathbb{R}$.

Hence letting $\theta \downarrow 0$ in results (4.40) and (4.41) it follows by the monotone convergence theorem that for $d_1, d_2 \in \mathbb{N}$ with $(d_1, d_2) \neq (0, 0)$,

$$ F_{ST}^{(0)}(d_1, d_2) = \frac{\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1 - \cos(d_1 x) \cos(d_2 y)}{M + \theta - M(\cos x + \cos y) / 2} \, dx \, dy}{8\pi^2 M + \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1 - \cos(d_1 x) \cos(d_2 y)}{M + \theta - M(\cos x + \cos y) / 2} \, dx \, dy} \tag{4.43} $$

$$ = \frac{\int_{0}^{2\pi} \frac{1 - \cos(d_1 x) (2 + 2\theta/M - \cos x - \sqrt{(2 + 2\theta/M - \cos x)^2 - 1})^2}{\sqrt{(2 + 2\theta/M - \cos x)^2 - 1}} \, dx}{\pi M + \int_{0}^{2\pi} \frac{1 - \cos(d_1 x) (2 + 2\theta/M - \cos x - \sqrt{(2 + 2\theta/M - \cos x)^2 - 1})^2}{\sqrt{(2 + 2\theta/M - \cos x)^2 - 1}} \, dx} \tag{4.44} $$

where we note that although the integrand in result (4.44) has a singularity at $x = 0$, its limit as $x \downarrow 0$ is finite, so that the integral in result (4.44) is finite (which implies that also the double integral in result (4.43) is finite). In the special case of $d_2 = 0$, Slatkin (1991) obtained expression (4.44) as an approximation valid for a stepping-stone model on a large square torus.
To find the value of $F_{ST}^{(1)}(d_1, d_2)$, we calculate

$$
\lim_{\theta \to 0} \frac{\partial}{\partial \theta} F_{ST}(d_1, d_2) = \pi M \lim_{\theta \to 0} \int_0^\pi \frac{1 - \cos(d_1z) \left( 2 + 2\theta/M - \cos z - \sqrt{(2 + 2\theta/M - \cos z)^2 - 1} \right)^{d_2}}{\sqrt{(2 + 2\theta/M - \cos z)^2 - 1}} \, dz
$$

$$
= \pi M \lim_{\theta \to 0} \int_0^\pi \frac{1 - \cos(d_1z) \left( 2 + 2\theta/M - \cos z - \sqrt{(2 + 2\theta/M - \cos z)^2 - 1} \right)^{d_2}}{\sqrt{(2 + 2\theta/M - \cos z)^2 - 1}} \, dz
$$

(4.45)

where we have applied Leibniz's rule (see for example Fikhtengol'ts 1965 p.141). The integral in the denominator of (4.45) is the same as that in (4.44) and is finite. Using (4.42) and Leibniz's rule, we have for $\theta > 0$

$$
\frac{\partial}{\partial \theta} \frac{1 - \cos(d_1z) \left( 2 + 2\theta/M - \cos z - \sqrt{(2 + 2\theta/M - \cos z)^2 - 1} \right)^{d_2}}{\sqrt{(2 + 2\theta/M - \cos z)^2 - 1}}
$$

$$
= \frac{M}{4\pi} \int_{-\pi}^\pi \frac{1 - \cos(d_1z) \cos(dy)}{\{M + \theta - M(\cos x + \cos y)/2\}^2} \, dy
$$

which is non-negative and monotonically increasing with decreasing $\theta$. By the monotone convergence theorem,

$$
\lim_{\theta \to 0} \int_0^\pi \frac{1 - \cos(d_1z) \left( 2 + 2\theta/M - \cos z - \sqrt{(2 + 2\theta/M - \cos z)^2 - 1} \right)^{d_2}}{\sqrt{(2 + 2\theta/M - \cos z)^2 - 1}} \, dz
$$

$$
= \int_0^\pi \lim_{\theta \to 0} \frac{1 - \cos(d_1z) \left( 2 + 2\theta/M - \cos z - \sqrt{(2 + 2\theta/M - \cos z)^2 - 1} \right)^{d_2}}{\sqrt{(2 + 2\theta/M - \cos z)^2 - 1}} \, dz.
$$

Calculating the derivative and the limit in the latter integral explicitly, a Taylor expansion of all cosines in the resulting expression for the integrand (which is continuous on $(0, \pi]$ but has a singularity at $x = 0$) shows that

$$
\lim_{\theta \to 0} \frac{1 - \cos(d_1z) \left( 2 + 2\theta/M - \cos z - \sqrt{(2 + 2\theta/M - \cos z)^2 - 1} \right)^{d_2}}{\sqrt{(2 + 2\theta/M - \cos z)^2 - 1}} = \frac{d_1^2 + d_2^2}{M}.
$$

Because

$$
\int_0^\pi \frac{1}{x} \, dx = +\infty
$$

it follows by the quotient test for improper integrals (Fikhtengol'ts 1965, p. 116) that for $d_1, d_2 \in \mathbb{N}$ with $(d_1, d_2) \neq (0, 0)$:

$$
\int_0^\pi \frac{1 - \cos(d_1z) \left( 2 + 2\theta/M - \cos z - \sqrt{(2 + 2\theta/M - \cos z)^2 - 1} \right)^{d_2}}{\sqrt{(2 + 2\theta/M - \cos z)^2 - 1}} \, dz = -\infty.
$$

Thus for $d_1, d_2 \in \mathbb{N}$ with $(d_1, d_2) \neq (0, 0)$

$$
\lim_{\theta \to 0} \frac{\partial}{\partial \theta} F_{ST}(d_1, d_2) = -\infty,
$$
Figure 4.13: The decrease of pairwise $F_{ST}$ values with increasing scaled migration rate, compared between the infinite stepping-stone models in one and two dimensions.

The thick lines are the values of $F_{ST}(100,0)$ (long-dashed line) and $F_{ST}^{(1)}(100,0)$ (solid line) of two colonies at distance (100,0) under a stepping-stone model on an infinite two-dimensional lattice. The thin lines are the corresponding values of a pair of colonies at distance 100 in an infinite line. For the exact $F_{ST}$ values, a mutation rate $\theta = 10^{-3}$ has been used. For both models, $M$ is the scaled total migration rate out of each colony.

as was also the case for the infinite one-dimensional stepping-stone model. Hence for $(d_1, d_2) \neq (0,0)$,

$$F_{ST}^{(1)}(d_1, d_2) = -\infty$$

and it cannot serve as an approximation for $F_{ST}(d_1, d_2)$.

All figures for this model have been obtained using the results in terms of single integrals, equations (4.41) and (4.44), applying an adaptive Newton-Cotes rule for numerical integration. Figure 4.13 shows the values of $F_{ST}(d_1, d_2)$ and $F_{ST}^{(1)}(d_1, d_2)$ for the infinite two-dimensional stepping-stone model (thick lines), plotted against the scaled migration rate $M$. Two colonies at distance (100,0) are considered. For comparison, the corresponding $F_{ST}$ values of two colonies 100 steps apart in the infinite one-dimensional stepping-stone model are also shown (thin lines). Much more even than for the finite models, the enormous difference between the homogenizing effects of one- and two-dimensional gene flow is noticed. In the two-dimensional model, an increasing level of gene flow rapidly decreases differences between the colonies, while in the one-dimensional model, the amount of subpopulation differentiation decreases only slowly with increasing migration rate, leaving very high $F_{ST}$ values even at large scaled migration rates. For a two-allele model, Kimura and Weiss (1964) and Slatkin (1985) observed that the amount of genetic variation between two colonies $d$ steps apart as measured by the probability of non-identity of two genes at distance $d$,
Figure 4.14: Pairwise $F_{ST}$ values as a function of the distance between the two colonies considered, compared between the infinite stepping-stone models in one and two dimensions. The filled symbols are the values of $F_{ST}(d, 0)$ (diamonds) and $F_{ST}^{(0)}(d, 0)$ (circles) of a pair of colonies in an infinite two-dimensional lattice. The symbols without fill indicate the corresponding values of a pair of colonies at distance $d$ in an infinite line. One symbol is plotted every 20th value of $d$. Scaled migration and mutation rates of respectively $M = 5$ and $\theta = 10^{-3}$ have been used.

$1 - f_d$ or $1 - f(d, 0)$, is much larger under the two-dimensional stepping-stone model than under the one-dimensional model (at all $d$, including $d = 0$). This is also the case for the mutation model assumed in this thesis. However, under two-dimensional dispersal the fraction $(1 - f_d)/(1 - \hat{f})$ (and hence the relative importance of variation within colonies) is larger, so that the relative importance of variation between the subpopulations, and hence the value of $F_{ST}$, is smaller than under the one-dimensional model (unless the two colonies are neighbouring, in which case the exact pairwise $F_{ST}$ value is slightly higher under the two-dimensional model than under the one-dimensional one, if the scaled migration rate $M$ is the same under both models). It is also seen that $F_{ST}^{(0)}$ is a better approximation for $F_{ST}$ in the case of an infinite two-dimensional lattice than it is for an infinite line.

In figure 4.14, the increase of genetic differentiation with increasing distance $d$ between two colonies in the infinite two-dimensional lattice (filled symbols) is compared to that in the infinite line (open symbols). It is seen that for the two-dimensional lattice, the increase of $F_{ST}(0, d)$ with $d$ is only important for small distances $d$ (smaller than 20, under the parameters of figure 4.14) and is much less substantial than under the infinite one-dimensional stepping-stone model. From equations (4.34) and (4.44), it can be seen that in both the infinite one- and two-dimensional stepping-stone models, the pairwise value of $F_{ST}^{(0)}$ converges to 1 as the distance tends to infinity,
while that of \( F_{ST} \) approaches a limiting value smaller than one. Unless \( M \) is very small, the limiting value of \( F_{ST} \) for the infinite two-dimensional lattice is much lower than that for the infinite line, giving for the infinite two-dimensional lattice a very large discrepancy between the values of \( F_{ST}(0, d) \) and \( F_{ST}^{(0)}(0, d) \), in the limit of \( d \to \infty \). Under the parameters of figure 4.14, the limiting value of \( F_{ST}(0, d) \) for the infinite two-dimensional lattice is 0.25, as \( d \to \infty \), yielding a limiting discrepancy of 300% between \( F_{ST}(0, d) \) and \( F_{ST}^{(0)}(0, d) \) (relative to the true \( F_{ST} \) value). However, whilst \( F_{ST} \) quickly approaches its limiting value, the increase of \( F_{ST}^{(1)} \) to 1 is extremely slow and these large discrepancies occur only for pairs of colonies an enormous distance apart (unlikely ever to be encountered in a real study).

### 4.2.6 The dependence of \( F_{ST} \) on the mutation rate

In this subsection the dependence of \( F_{ST} \) on the scaled mutation rate \( \theta \) and, related to this, the accuracy of both approximations \( F_{ST}^{(0)} \) and \( F_{ST}^{(1)} \), are more closely examined and compared between the different models of population structure studied in the previous subsections. Because \( F_{ST} \) values for pairs of colonies (as calculated for the stepping-stone models) are systematically lower than global \( F_{ST} \) values, the global \( F_{ST} \) values for the symmetric island model can be compared only with the global \( F_{ST} \) values of the (finite) stepping-stone models, given by equations (4.30) to (4.32) and (4.37) to (4.39).

In figure 4.15, the global values of \( F_{ST} \) (thick line), Slatkin's approximation \( F_{ST}^{(0)} \) (thin solid line) and the first order approximation \( F_{ST}^{(1)} \) (dashed line) are plotted against the scaled mutation rate \( \theta \), for a symmetric island model (figure 4.15a), a square torus (figure 4.15b), a narrow torus (figure 4.15c) and a circle (figure 4.15d). In all cases the population consists of 100 subpopulations and the total scaled migration rate out of each colony is \( M = 5 \). It seems a general fact that the amount of subpopulation differentiation (as measured by \( F_{ST} \)) decreases with increasing mutation rate \( \theta \). This means that although the amount of genetic variation in the population, as measured by \( 1 - \bar{f} \), increases with increasing mutation rate, the fraction \( (1 - \bar{f}_0)/(1 - \bar{f}) \) increases, so that \( F_{ST} \) decreases. Slatkin's approximation is an upper bound for \( F_{ST} \), while the first order approximation yields a lower bound. Comparing the different models in figure 4.15, one observes once more the enormous increase of \( F_{ST} \) as gene flow becomes increasingly restricted to one dimension. There is least subpopulation differentiation under the symmetric island model, where gene dispersal is to all other colonies. There is more differentiation in the torus, where migration is restricted to the four neighbouring colonies. The value of \( F_{ST} \) increases from a square 10 x 10 torus to a narrow 25 x 4 torus, and is largest under the one-dimensional circular stepping-stone model, where gene flow is most restricted. Furthermore, going from figure 4.15a to figure 4.15d (from the island model, via the toruses, to the circle), the derivative of \( F_{ST} \) at \( \theta = 0 \) becomes increasingly negative. Whereas under the symmetric island model, \( F_{ST} \) depends only very weakly on the
Figure 4.15: $F_{ST}$ and its approximations as a function of the scaled mutation rate, compared between different models of population structure.

This figure shows the global values of $F_{ST}$ and both its approximations for a symmetric island model (figure a), a stepping-stone model on a $10 \times 10$ torus (figure b), a stepping-stone model on a $25 \times 4$ torus (figure c) and a circular stepping-stone model (figure d). For all models we assumed 100 subpopulations and migration rate $M = 5$. In all plots, the thick solid line is the exact $F_{ST}$ value, the thin solid line represents Slatkin's approximation $F^{(0)}_{ST}$, and the long-dashed line shows the first order approximation $F^{(1)}_{ST}$. 

\begin{align*}
\text{a) Symmetric island model} \\
\text{b) Stepping-stone model on a } 10 \times 10 \text{ torus} \\
\text{c) Stepping-stone model on a } 25 \times 4 \text{ torus} \\
\text{d) Circular stepping-stone model}
\end{align*}
mutation rate, the dependence of $F_{ST}$ on $\theta$ becomes more important as gene dispersal is more restricted. Under the one-dimensional stepping-stone model (figure 4.15d), $F_{ST}$ depends strongly on the scaled mutation rate, even for small mutation rates. Under the parameters of figure 4.15, $F_{ST}$ decreases under the symmetric island model by only 14% as $\theta$ increases from 0 to 1; under the circular stepping-stone model (figure 4.15d), $F_{ST}$ falls off by 70% as $\theta$ increases from 0 to 1 and by almost 40% as $\theta$ increases from 0 to 0.1. As a consequence of the increasing dependence of $F_{ST}$ on the mutation rate, the accuracy of both Slatkin's and the first order approximation decreases markedly from the island model to the circle. While it is seen that both approximations are very accurate for the island model even for quite high scaled mutation rates, when dispersal is very restricted they rapidly become highly inaccurate with increasing scaled mutation rate and some care should be taken when using these approximations. Under the parameters of figure 4.15d, neither approximation should be used for $\theta$-values of order larger than $10^{-2}$. For $\theta = 10^{-2}$, the values of $F_{ST}$, $F_{ST}^{(0)}$ and $F_{ST}^{(1)}$ are 0.69, 0.77 and 0.65, respectively. For $\theta = 0.1$ (for example, a Wright-Fisher model with $N = 10000$ and $\mu = 2.5 \times 10^{-6}$), the value of $F_{ST}^{(0)}$ is 64% higher than the real $F_{ST}$ value (0.77 against 0.47), while the value of $F_{ST}^{(1)}$ is negative.

In figure 4.16 we illustrate how $F_{ST}$ values of pairs of colonies depend on the scaled mutation rate, for the finite and infinite stepping-stone models in one and two dimensions. It is seen at once that this dependence on $\theta$ is very similar to that of the global $F_{ST}$ values in figure 4.15. The values of $F_{ST}(d)$ ($F_{ST}(0,d)$ for the two-dimensional models) and its approximations are plotted for two colonies at distance $d = 100$ on a circle with 200 subpopulations (figure 4.16a), an infinite line (figure 4.16b), a $200 \times 200$ torus (figure 4.16c) and an infinite two-dimensional lattice (figure 4.16d). Under the stepping-stone models, $F_{ST}$ depends more strongly on the mutation rate as there are more subpopulations. (As a consequence, both Slatkin's and the linear approximation lose accuracy as the number of subpopulations increases, as was seen in figures 4.6 and 4.12.) The dependence of the pairwise value of $F_{ST}$ on the mutation rate is maximal under the infinite stepping-stone models in one and two dimensions, where the derivative of $F_{ST}(d)$ and $F_{ST}(0,d)$ with respect to $\theta$, at $\theta = 0$, is $-\infty$. Under the infinite one-dimensional stepping-stone model, a scaled mutation rate of $\theta = 36 \times 10^{-3}$ (for example, groups of $N = 9000$ diploids and mutation rate $\mu = 10^{-6}$ under a Wright-Fisher model) suffices (under the parameters of figure 4.16) to reduce the value of $F_{ST}(d)$ to only half the value of its limit as $\theta \downarrow 0$. Although under the two-dimensional stepping-stone model the decrease of the pairwise $F_{ST}$ value with increasing mutation rate is less substantial than under the one-dimensional model, it may still be important if the number of colonies is very large. Under the parameters of figure 4.16d and with $\theta = 36 \times 10^{-3}$ as above, the value of $F_{ST}(0,d)$ is 64% of the value of $F_{ST}^{(0)}(0,d)$ (0.18 against 0.28).

The rapid decrease of $F_{ST}$ (or $F_{ST}(d)$) with increasing scaled mutation rate $\theta$, observed most distinctly for the one-dimensional stepping-stone model with a large number of colonies, contradicts
Chapter 4: SUBPOPULATION DIFFERENTIATION

Figure 4.16: Pairwise $F_{ST}$ values as a function of the scaled mutation rate, compared between the finite and infinite stepping-stone models in one and two dimensions.

Pairwise values of $F_{ST}$ and its approximations are shown under a circular stepping-stone model with 200 subpopulations (figure a), an infinite linear stepping-stone model (figure b), a stepping-stone model on a $200 \times 200$ torus (figure c) and an infinite two-dimensional stepping-stone model (figure d). The distance between the two colonies considered is $d = 100$ in the one-dimensional stepping-stone models and $(d_1, d_2) = (0, 100)$ in the two-dimensional stepping-stone models. The thick solid line is the pairwise value of $F_{ST}$, the thin lines those of its approximations $F^{(0)}_{ST}$ (solid line) and $F^{(1)}_{ST}$ (long-dashed line). For the infinite stepping-stone models (figures b and d), the first order approximation $F^{(1)}_{ST}$ could not be shown, as its value is $-\infty$. In all four models, a scaled migration rate of $M = 5$ was assumed.
the common belief that $F_{ST}$ would be nearly independent of the mutation rate when the mutation rate is small. The derivative of $-\infty$ at $\theta = 0$, found for the infinite stepping-stone models in one and two dimensions, implies precisely that for small mutation rates, the dependence of $F_{ST}$ on $\theta$ (and hence on $\mu$) under these models could not possibly be any stronger. Nevertheless, mutation rates at loci coding for electrophoretic enzymes can be so small that even under very restricted dispersal it may not make much difference if the $F_{ST}$ value is calculated assuming too small a mutation rate (although care should be taken). With the parameters used for the figures in this subsection, the mutation rate $\theta$ should be taken into account when it is of order $10^{-2}$ or larger and dispersal is restricted. (Even smaller mutation rates may have an important effect on the $F_{ST}$ value of a pair of subpopulations, compared to the limit of zero mutation rate, if their distance is much larger than in figure 4.16.) If it is known that the structure of the population resembles that of a symmetric island model or the population occupies a truly two-dimensional habitat with not too large a number of colonies, the mutation rate seems to affect the value of $F_{ST}$ only weakly as long as the scaled mutation rate $\theta$ is of order smaller than one. A scaled mutation rate $\theta$ of order one or larger (which may occur, for example, at VNTR loci) is likely to cause an important reduction of $F_{ST}$, compared to $F_{ST}$ values for small scaled mutation rates.

While we have observed that $F_{ST}$ strongly depends on the details of population structure for the range of mutation rates shown in figures 4.15 and 4.16, this is no longer true, for the models studied, for very large scaled mutation rates (which may still be realistic for VNTR loci). This is illustrated in figure 4.17, where we plot the exact values of $F_{ST}$ under the same models and parameters as used in figure 4.15, but this time for a very large range of the scaled mutation rate $\theta$. For the parameters of this figure, $F_{ST}$ is approximately independent of the precise population structure when the scaled mutation rate $\theta$ is of order 10 or larger. We have checked that the $F_{ST}$ values of the island and stepping-stone models are always very close to each other when $\theta$ is at least an order of magnitude larger than $M$. In fact, denoting $U = n\theta$ (which is, under the Wright-Fisher model for reproduction within colonies, twice the number of mutations expected in the total population per generation), we have under both the island and stepping-stone models that

$$\lim_{\theta \to \infty} \frac{F_{ST}}{1 + U/(n-1)} = 1.$$  \hspace{1cm} (4.46)

This asymptotic behaviour of $F_{ST}$ for very large mutation rates is established under more general conditions in Chapter 5 (Theorem 5.2 and Corollary 5.3). For the pairwise $F_{ST}$ values under the symmetric island and stepping-stone models,

$$\lim_{\theta \to \infty} \frac{F_{ST}(i,j)}{1 + 2\theta} = 1,$$

where $F_{ST}(i,j)$ denotes the $F_{ST}$ value of the pair of colonies $i$ and $j$. 

---

**Chapter 4: SUBPOPULATION DIFFERENTIATION**

103
Figure 4.17: Global $F_{ST}$ values as a function of the scaled mutation rate, for different models of population structure.

The global value of $F_{ST}$ is shown for a symmetric island model (solid line), a 10 x 10 torus (dotted), a 25 x 4 torus (long-dashed) and a circular stepping-stone model (dot-dashed). The circle, filled diamond, open diamond and square on the vertical axis represent the corresponding $F_{ST}^{(0)}$ values. For all models we assumed 100 subpopulations and a scaled migration rate of $M = 5$. 

Chapter 4: SUBPOPULATION DIFFERENTIATION

4.2.7 Discussion

We have shown that the decrease of $F_{ST}$ with increasing mutation rate may be more important than is generally believed, particularly when gene flow is restricted to one dimension. The figures indicate that Slatkin's approximation $F_{ST}^{(0)}$ will be an accurate approximation for $F_{ST}$ under most realistic conditions. However, a substantial discrepancy between $F_{ST}^{(0)}$ and the exact $F_{ST}$ value may occur in the case of highly mutable loci and/or a narrow habitat (dispersal restricted to essentially one dimension). Although it seems hard to prove this for an arbitrary model of population structure, the various figures showing the dependence of $F_{ST}$ on the mutation rate suggest as a general fact that $F_{ST}$ is a monotonically decreasing and convex function of the mutation rate. As a consequence, Slatkin's approximation $F_{ST}^{(0)}$, which is the limit of $F_{ST}$ as the mutation rate tends to zero, yields an upper bound on $F_{ST}$, while the first order approximation $F_{ST}^{(1)}$ is a lower bound for $F_{ST}$.

Among the models studied in this section, the decrease of $F_{ST}$ with increasing mutation rate, and hence the discrepancy between $F_{ST}$ and its approximation $F_{ST}^{(0)}$, was (for a fixed number of subpopulations) largest under a one-dimensional stepping-stone model, where gene dispersal is most restricted. Upper bounds on the discrepancy between $F_{ST}^{(0)}$ and the exact $F_{ST}$ value under a circular stepping-stone model are readily derived from equations (4.31), (4.32), (4.25) and (4.26). For the global $F_{ST}$ values, we have for the absolute error:

$$ F_{ST}^{(0)} - F_{ST} \leq \frac{F_{ST}^{(0)} - F_{ST}^{(1)}}{F_{ST}^{(0)}} $$

$$ = \frac{(n\theta)}{5} \cdot \frac{n^4 + 10n^2 - 11}{(n^2 - 1 + 6nM)^2} $$

$$ \leq \frac{n\theta}{(1 + \frac{6M}{n})^2} \quad (4.47) $$

and for the error$^6$ relative to $F_{ST}^{(0)}$:

$$ \frac{F_{ST}^{(0)} - F_{ST}}{F_{ST}^{(0)}} \leq \frac{F_{ST}^{(0)} - F_{ST}^{(1)}}{F_{ST}^{(0)}} $$

$$ = \frac{(n\theta)}{5} \cdot \frac{n^4 + 10n^2 - 11}{(n^2 - 1)(n^2 - 1 + 6nM)} $$

$$ \leq \frac{n\theta}{1 + \frac{6M}{n}} \quad (4.48) $$

($n \geq 2$). For the pairwise $F_{ST}$ values, it follows similarly that

$$ F_{ST}^{(0)}(d) - F_{ST}(d) \leq \frac{(n\theta)}{3} \cdot \frac{1 + \frac{2}{d(n-d)}}{(1 + \frac{2M}{d(n-d)})^2} $$

$$ \leq \frac{n\theta}{(1 + \frac{2M}{d})^2} $$

$^6$Where before, we measured the error relative to the exact $F_{ST}$ value, it is more convenient here to measure it relative to $F_{ST}^{(0)}$. 


and
\[
\frac{F_{ST}^{(0)}(d) - F_{ST}(d)}{F_{ST}^{(0)}(d)} \leq \frac{(n\theta)}{3} \cdot \frac{1 + \frac{2}{d(n-d)}}{1 + \frac{2M}{d(n-d)}}
\]
\[
\leq \frac{n\theta}{1 + \frac{2M}{d}}.
\]

Hence, under the circular stepping-stone model, Slatkin's approximation for $F_{ST}$ is accurate whenever $(n\theta)$ is small relative to $\max(1, M/n)$, in the case of global $F_{ST}$ values, and whenever $(n\theta)$ is small relative to $\max(1, M/d)$, in the case of pairwise $F_{ST}$ values. Note that these criteria are in terms of the absolute number of mutations occurring in the total population per generation.

They are consistent with the observation made earlier that under a one-dimensional stepping-stone model with many colonies, a large discrepancy between $F_{ST}$ and $F_{ST}^{(0)}$ can still occur when $M$ is a few orders of magnitude larger than $\theta$.

Because among the models studied in this section, the one-dimensional stepping-stone models are the ones for which (for a fixed number of subpopulations) Slatkin's approximation $F_{ST}^{(0)}$ is least accurate, we believe the error bounds and criteria for accuracy given above to be valid also under the two-dimensional stepping-stone model (replacing $d$ by $d_1 + d_2$ in the error bounds and criteria) and the symmetric island model. As one might think of gene flow being maximally restricted under the one-dimensional stepping-stone models, one might expect these bounds and criteria to be valid under any model of population structure. However, we will show in the next section that this is not true, as in less symmetric population structures the decrease of $F_{ST}$ with increasing mutation rate can be even more rapid than under the (finite) one-dimensional stepping-stone model.

For many population structures, the above criteria (based on a situation in which gene dispersal is very restricted) will be much stronger than necessary. For example, under the symmetric island model, it follows from equations (4.16) and (4.17) that
\[
\frac{F_{ST}^{(0)} - F_{ST}}{F_{ST}^{(0)}} \leq \frac{2\theta}{1 + M}.
\]

Hence, when the structure of the population is well described by the symmetric island model, $F_{ST}$ will be close to Slatkin's approximation $F_{ST}^{(0)}$ as soon as $\theta$ is small relative to $\max(1, M)$. Note that the latter criterion is in terms of the number of mutations per colony per generation, and does not depend on the number of colonies.

The above criteria for $F_{ST}^{(0)}$ to be a good approximation for $F_{ST}$ all involve the migration rate $M$. Since in real studies $M$ is usually unknown and often to be estimated from the $F_{ST}$ value, the following bound on the discrepancy between the global values of $F_{ST}^{(0)}$ and $F_{ST}$ may be of more practical interest: for both the symmetric island model and the finite stepping-stone models in one and two dimensions, the global $F_{ST}$ values satisfy
\[
F_{ST}^{(0)} - F_{ST} \leq U \left(F_{ST}^{(0)}\right)^2,
\] (4.49)
where \( U = n \theta \). Hence the error relative to \( F_{ST}^{(0)} \) is subject to

\[
\frac{F_{ST}^{(0)} - F_{ST}}{F_{ST}^{(0)}} \leq UF_{ST}^{(0)},
\]

so that \( F_{ST}^{(0)} \) is an accurate approximation for \( F_{ST} \) whenever \( \left( UF_{ST}^{(0)} \right) \) is small, which is a criterion in terms of the number of mutations occurring per generation in the total population and the value of Slatkin's approximation \( F_{ST}^{(0)} \). For the pairwise \( F_{ST} \) values, it is easily seen that the inequalities (4.49) and (4.50) are valid under the circular stepping-stone model and under the symmetric island model. For the stepping-stone model on the torus, we have checked (4.49) and (4.50) for a very large number of parameter values: these inequalities are true (and in fact less tight than under the circular stepping-stone model) for all the values checked, yet they have not been proved analytically for the torus.

The various bounds we have derived above on the error made by using \( F_{ST}^{(0)} \) as an approximation for \( F_{ST} \) also give an idea of how much the \( F_{ST} \) value is affected if an incorrect value of the scaled mutation rate is used. Because for the models studied, \( F_{ST} \) is monotonically decreasing and convex in the mutation rate, the difference between the exact \( F_{ST} \) values for two different mutation rates \( \theta_1 \) and \( \theta_2 \) is subject to

\[
|F_{ST}(\theta_1) - F_{ST}(\theta_2)| \leq -|\theta_1 - \theta_2| \left| \lim_{\theta \to 0} \frac{\partial}{\partial \theta} F_{ST} \right| = |F_{ST}^{(1)}(\theta_1) - F_{ST}^{(1)}(\theta_2)|.
\]

From this and from the analytical results obtained in the previous subsections it follows that all the bounds derived in this subsection on the absolute and relative discrepancies between \( F_{ST}^{(0)} \) and \( F_{ST} \) change into upper bounds on, respectively, the absolute difference between \( F_{ST}(\theta_1) \) and \( F_{ST}(\theta_2) \) and their difference relative to \( F_{ST}^{(0)} \), if in the inequalities in this subsection, one replaces \( (F_{ST}^{(0)} - F_{ST}) \) by \( |F_{ST}(\theta_1) - F_{ST}(\theta_2)| \) in the left-hand side, while \( \theta \) is replaced by \( |\theta_1 - \theta_2| \) in the right-hand side. In particular, the global \( F_{ST} \) values of both the symmetric island model and the finite stepping-stone models satisfy

\[
|F_{ST}(\theta_1) - F_{ST}(\theta_2)| \leq n|\theta_1 - \theta_2| \left( F_{ST}^{(0)} \right)^2.
\]

In this way, the various criteria for \( F_{ST}^{(0)} \) to be an accurate approximation for \( F_{ST} \) give rise to criteria under which (assuming \( \theta_1 \) and \( \theta_2 \) are an upper and a lower bound on the real mutation rate \( \theta \)) the use of an incorrect mutation rate (in the range \( \theta_1 \) to \( \theta_2 \)) has negligible effect on the \( F_{ST} \) value.

The figures shown for the stepping-stone models indicate that the first order approximation \( F_{ST}^{(1)} \) in itself is, in the absence of information on the population structure, too unreliable to be used as an approximation for \( F_{ST} \) in practical studies. Nevertheless the value of \( F_{ST}^{(1)} \) may be useful for assessing the accuracy of Slatkin's approximation \( F_{ST}^{(0)} \). If it is possible to obtain reliable estimates of both \( F_{ST}^{(0)} \) and \( F_{ST}^{(1)} \), then, because of the conjectured monotonic decreasing and convex character
Chapter 4: SUBPOPULATION DIFFERENTIATION

of $F_{ST}$ as a function of $\theta$, the real $F_{ST}$ value would be expected to lie in between these estimates (and the difference between these estimates yields an estimated upper bound on the error made by using Slatkin’s approximation for $F_{ST}$). Assuming the infinite sites model\(^7\) of neutral mutation (which can be regarded as a specific and more detailed case of the infinite alleles model) and denoting by $S_0$ the number of nucleotide differences between a pair of genes sampled at random from a single subpopulation and by $S$ that between two genes sampled from the collection of subpopulations considered, the two approximations for $F_{ST}$ can be expressed as

$$F_{ST}^{(0)} = \frac{ES - ES_0}{ES}$$

and

$$F_{ST}^{(1)} = F_{ST}^{(0)} - \frac{E(S^2)ES_0 - E(S)E(S)}{2(ES)^2}$$

(provided $S$ and $S_0$ have finite second moments), where we have used equations (4.8) and (4.10) and the fact that, given the coalescence time $t$ of a pair of genes, their number of nucleotide differences is Poisson distributed with mean $\theta t$ (where $\theta$ is the scaled mutation rate and time-scaling is as before). This implies that the values of both Slatkin’s and the first order approximation for $F_{ST}$ could be estimated from DNA sequence data by counting the numbers of nucleotide differences between pairs of sampled genes. However, in any such estimation procedure, one should keep in mind the uncertainties and possible biases arising from correlations between the numbers of nucleotide differences of even disjoint pairs of genes, induced by the common genealogical tree of the population.

The coefficient $F_{ST}$ is mainly used to estimate levels of gene flow between subpopulations, assuming a symmetric island model of population structure with a large number of subpopulations and a small mutation rate. However, we have shown that the rate of decrease of $F_{ST}$ with increasing migration rate varies importantly between different population structures. It would therefore be of interest to investigate explicitly the impact of the actual population structure and mutation rate on the accuracy of these estimated levels of gene flow. We also note that it may be possible to use the derivative of $F_{ST}$ with respect to the mutation rate at mutation rate zero (which may be estimated from DNA sequence data) to infer information on the underlying population structure.

4.3 Some less symmetric structures

To our knowledge, nearly all published theoretical studies of genetic differentiation have been restricted to population structures in which all subpopulations are identical with respect to size

---

\(^7\)Under the infinite sites model, the locus considered is composed of an infinite number of nucleotide sites, at each of which a mutation can occur. The number of mutating sites per gene per generation is assumed to be Poisson distributed with mean $\mu$. Because of the assumption of an infinite number of sites it follows that no two mutations will ever affect the same site.
and migration pattern. Most analyses assume also that all non-zero migration rates are equal (a few exceptions are studies by Kimura and Weiss 1964, Maruyama 1970 and Malécot 1975 of genetic variability in two- and three-dimensional stepping-stone models with different migration rates in the different dimensions of the model, and studies by Nagylaki 1988 and Nagylaki and Barcilon 1988 of the probability of identity of a pair of genes in an infinite linear stepping-stone model with spatial variation in the migration rates or with a geographic barrier). For many natural populations these assumptions are, of course, unrealistic. In this section we examine how asymmetries in the population structure may affect the amount of subpopulation differentiation, as measured by $F_{ST}$. In Subsections 1 to 3, the values of $F_{ST}$ and its approximations are calculated explicitly for the three models of population structure introduced in Section 3.3. Figures illustrate for each model how $F_{ST}$ depends on the different parameters of the model, where attention is paid particularly to the dependence of $F_{ST}$ on the level of asymmetry in the structure. In Subsection 4, our observations are discussed and compared with some of the observations made for the symmetrical models in the previous section.

We will in this section be concerned only with the global values of $F_{ST}$ and its approximations. In Subsection 4.1.2, $F_{ST}$ was expressed in exact and approximate ways in terms of coalescence times of pairs of genes sampled from the population. When a gene is sampled from the population and subpopulation sizes are unequal, there seem two distinct, natural, sampling schemes. If the population subdivision is geographic and genes are sampled at specified locations, it is known which subpopulation each gene is taken from, but the relative sizes of the different subpopulations are, in most cases, unknown. In those cases, equal weighting of the subpopulations may be appropriate, that is, to choose a gene at random from the population, we first choose one of the subpopulations, with each choice being equally likely, regardless of its size, and then choose a gene uniformly at random from the chosen subpopulation. Assuming that the population consists of $n$ subpopulations (where $n$ is finite) and labelling the subpopulations as $1, 2, \ldots, n$, the probabilities of identity, $f_0$ and $\bar{f}$, of, respectively, two genes sampled from a single subpopulation and two genes sampled from the total population are in this context given by

\[
\begin{align*}
  f_0 &= \frac{1}{n} \sum_{i=1}^{n} f_{ii} \\
  \bar{f} &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij},
\end{align*}
\]

(4.51)

where $f_{ij}$ is the probability of identity of a gene from subpopulation $i$ and a gene from subpopulation $j$. In other settings (perhaps for example human populations), one may suspect population subdivision, but it may be difficult to outline the different subpopulations and one may have a random sample from the population, not knowing which individuals belong to which subpopulations. In such cases, it may be realistic to assume that all individuals in the population are
equally likely to be sampled, so that the various subpopulations are naturally weighted by their relative sizes:

\[
\begin{align*}
    f_0 &= \sum_{i=1}^{n} P_i f_{ii} \\
    \bar{f} &= \sum_{i=1}^{n} \sum_{j=1}^{n} P_i P_j f_{ij},
\end{align*}
\]

where \( P_i \) is the proportion of the total population that belongs to subpopulation \( i \). Other weightings, for example weighting by sample size (Weir and Cockerham 1984), seem less appropriate for a theoretical analysis of the kind performed in this thesis. For many purposes, equal weighting of subpopulations seems a natural choice and this will be our primary focus in this section.

As in the previous sections, \( \theta \) denotes the scaled mutation rate (every gene mutates at scaled rate \( \theta/2 \)). If the size of subpopulation \( i \) is \( 2c_i N \) genes \((i = 1, \ldots, n)\) and reproduction within each subpopulation is according to the neutral Wright-Fisher model, time-scaling is in units of \( 2N \) generations and \( \theta = \lim_{N \to \infty} (4N\mu) \), where \( \mu \) is the probability of mutation per gene per generation. In that case, \( c_i \theta \) is (in the coalescent approximation) twice the number of mutations expected in subpopulation \( i \) per generation. We also denote

\[
U := c\theta
\]

(recall that \( c = \sum_{i=1}^{n} c_i \)), which will also be referred to as the "mutation rate". Under the Wright-Fisher model, \( U \) is twice the number of mutations expected in the total population per generation.

### 4.3.1 The general two-population model

We recall the definition of the "migration rate" \( M \) for this model (see equation (3.55)). In the case of the Wright-Fisher model, \( M \) is (in the coalescent approximation) twice the number of genes exchanged between the two subpopulations each generation. Our analysis is in terms of the parameters \( M, U, c \) and \( P \), the proportion of genes in subpopulation 1.

The (global) value of \( F_{ST} \) under equal weighting of the subpopulations is calculated from equation (4.3), where \( f_0 \) and \( \bar{f} \) are calculated according to equations (4.51) from the results for the moment-generating function of the coalescence time of a pair of genes, equations (3.56) to (3.58) (recall that the probability of identity by descent of a pair of genes is equal to the moment-generating function of their coalescence time, evaluated in \( s = \theta \)). We find:

\[
F_{ST} = \frac{M + 2P(1-P)(2+U)}{M(3+4M) + P(1-P)(4 + 6U^2 + 8P(1-P)(2+U^2)}.
\]

To calculate the value of Slatkin's approximation \( F_{ST}^{(0)} \) under equal weighting, the mean coalescence time of two genes sampled from a single subpopulation, \( E T_0 \), and that of two genes sampled
from the total population, $ET$, are obtained from equations (3.59) to (3.61) by taking averages analogous to (4.51). Whereas under the symmetric models studied in the previous section, the mean coalescence time of two genes from a single subpopulation is independent of the migration rate, we found in Section 3.3.1 that $ET_{11}$ and $ET_{22}$ under the general two-population model do depend on the migration rate, unless $P = 1/2$. Note in particular that under equal weighting, $ET_0 = (ET_{11} + ET_{22})/2$ also depends on the migration rate. Equation (4.8) gives

$$F_{ST}^{(0)} = \frac{4P(1 - P) + M}{4P(1 - P)(1 + 2M) + M(3 + 4M)}.$$  

Alternatively, this result could have been obtained from (4.54) according to the definition of $F$, equation (4.7).

The value of the first order approximation $F_{ST}^{(1)}$ under equal weighting is found from result (4.54) according to equation (4.9), or from equation (4.10), where $E[T_3^2]$ and $E[T_4^2]$ are calculated from results (3.62) to (3.64) analogously to equations (4.51):

$$F_{ST}^{(1)} = F_{ST}^{(0)} - 4P(1 - P)U \frac{4P(1 - P)(1 + 2M) + M^2}{4P(1 - P)(1 + 2M) + M(3 + 4M)}.$$

Figure 4.18 shows the values of $F_{ST}$ and its approximation $F_{ST}^{(0)}$ under equal weighting of the subpopulations, plotted as a function of the proportion $P$ of genes in colony 1, for two different values of the mutation rate and for two different levels of gene flow. Because there appears to be some confusion in the population genetics literature about the generality under which $ET_0 = c$ (for example, Slatkin 1993), we have also plotted the values (dotted lines) that would be obtained for $F_{ST}^{(0)}$ if one incorrectly assumed that $ET_0 = c$. Clearly, this value is not a good approximation for $F_{ST}$ if the colony sizes are unequal. It is seen that under equal weighting, $F_{ST}$ depends less on the relative colony sizes than one might have expected. Unless the mutation rate is very high or the subpopulations are very near complete isolation, the $F_{ST}$ value is maximal when the two subpopulations are of equal size and decreases as the model is increasingly asymmetric. Under those conditions, both the amount of genetic variation in the total population (as measured by $1 - \bar{f}$) and the average amount of genetic variation within the subpopulations $(1 - f_0)$ are maximal in the symmetric model; as $P$ deviates further from 1/2, the amount of variation between colonies is affected more than that within colonies, so that $F_{ST}$ decreases. The dependence of $F_{ST}$ on the relative colony sizes is seen to be more important when there is fairly little gene flow between the two colonies (at larger values of $M$, the plot of $F_{ST}$ against $P$ is almost a straight line) and, for the range of mutation rates shown, when the mutation rate is very small ($F_{ST}^{(0)}$ depends more on $P$ than $F_{ST}$). Finally, it is observed that the approximation $F_{ST}^{(0)}$ (solid lines) is more accurate as the model is more asymmetric.

Figure 4.19 shows the decrease of $F_{ST}$ (under equal weighting) with increasing level of gene flow, for a symmetric two-population model (dotted line), a model in which nine tenths of the
Figure 4.18: $F_{ST}$ values as a function of the relative subpopulation sizes in the general two-population model, under equal weighting of the subpopulations.

The $F_{ST}$ values under equal weighting are shown for two different values of the "migration rate" $M$: the thick lines are for $M = 1$, for the thin lines we assumed $M = 5$. The dot-dashed and long-dashed lines are the exact $F_{ST}$ values for, respectively, $U = 1$ and $U = 0.1$. The solid lines represent the approximation $F_{ST}^{(0)}$. The dotted lines indicate the value one would obtain for $F_{ST}^{(0)}$ falsely assuming $E_{T0} = c$. 
Figure 4.19: \( F_{ST} \) values as a function of the scaled migration rate, compared between symmetric and asymmetric two-population models, under equal weighting of the subpopulations.

This figure shows the exact \( F_{ST} \) values of three two-population models, with respectively \( P = 1/2 \) (dotted line), \( P = 1/10 \) (long-dashed) and \( P = 1/100 \) (dot-dashed). For all three models, we assumed \( U = 0.1 \). The solid line represents Slatkin's approximation \( F^{(S)}_{ST} \) for the symmetric model \( (P = 1/2) \).

population belong to one colony and one tenth to the other (long-dashed line), and one in which 99% of the genes belong to the same colony (dot-dashed line). The limiting value of \( F_{ST} \) as the colonies approach complete isolation,

\[
\lim_{M \to 0} F_{ST} = \frac{2 + U}{2 + 3U + 4P(1 - P)U^2},
\]

is nearly independent of \( P \) if \( U \) is small. Surprisingly, this limiting \( F_{ST} \) value is higher as the model is more asymmetric: under almost complete isolation, unequal colony sizes can enhance subpopulation differentiation (as measured by \( F_{ST} \)). An increasing level of gene flow has a stronger homogenizing effect on the genetic composition of the population (\( F_{ST} \) values decrease faster with increasing \( M \)) as the colony sizes are more unequal. Because in practical studies the subpopulation sizes are usually unknown, one might want to approximate \( F_{ST} \) by the \( F^{(S)}_{ST} \) value of the symmetric model (solid line). In doing so, one does not make a big error, provided there is a fair amount of gene flow between the two colonies and the mutation rate is not too large. Note however that, if the two subpopulations are quite isolated (\( M \) smaller than 1, that is under the Wright-Fisher model, less than one gene exchanged per two generations) and have very unequal sizes, the discrepancy with the real \( F_{ST} \) value can be large: about \( M = 0.1 \), the \( F^{(S)}_{ST} \) value of the symmetric model is twice the \( F_{ST} \) value of the model with \( P = 1/100 \). As the \( F^{(S)}_{ST} \) value for \( P = 1/100 \) (not plotted) practically coincides with the \( F_{ST} \) value shown (dot-dashed line), this large discrepancy still occurs.
Chapter 4: SUBPOPULATION DIFFERENTIATION

Figure 4.20: $F_{ST}$ and its approximations as a function of the scaled mutation rate, compared between a symmetric and an asymmetric two-population model (equal weighting).

The thin lines are the values of $F_{ST}$ and its approximations for a symmetric two-population model ($P = 1/2$), the thick lines those of a two-population model with $P = 1/20$. The exact $F_{ST}$ values are indicated by a solid line, the approximations $F_{ST}^{(0)}$ and $F_{ST}^{(1)}$ by dotted and long-dashed lines, respectively. The values are plotted against $U = c\theta$. A migration rate of $M = 1$ is assumed.

for smaller mutation rates as well.

In figure 4.20, the values of $F_{ST}$ and both its approximations are shown as a function of the "mutation rate" $U$, for $P = 1/2$ (thin lines) and for $P = 1/20$ (thick lines). It is observed that the decrease of $F_{ST}$ with increasing mutation rate is much weaker in the asymmetric model than in the symmetric one. As a consequence, both approximations $F_{ST}^{(0)}$ and $F_{ST}^{(1)}$ are highly accurate for the asymmetric two-population models, unless the mutation rate is very high. One also sees that at very large mutation rates (for example, at VNTR loci), $F_{ST}$ under equal weighting depends strongly on the relative colony sizes. From equation (4.54) it can in fact be seen that in the limit as $\theta$ tends to infinity, the $F_{ST}$ value under equal weighting of a two-population model in which a proportion $P$ of the genes live in colony 1 is a factor $1/[4P(1 - P)]$ larger than the $F_{ST}$ value of the symmetric model. Hence, in contrast with the observations made for the symmetric island and stepping-stone models in the previous section, $F_{ST}$ under equal weighting in the two-population model still depends on the details of population structure when the mutation rate is very large.

The (global) value of $F_{ST}$ under weighting by size is calculated from equation (4.3), where $f_0$ and $\bar{f}$ are calculated according to equations (4.52) from results (3.56) to (3.58). We find:

$$F_{ST} = \frac{P(1 - P)(M + 4P(1 - P) + 2P(1 - P)U)}{P(1 - P)(M + 2U + P(1 - P)(4 - 2U + 2U^2) + 3MU) + M(1 + M)}. \quad (4.55)$$
Chapter 4: SUBPOPULATION DIFFERENTIATION

The value of Slatkin’s approximation $F_{ST}^{(0)}$ under weighting by size is found by letting $\theta$ (and hence $U$) tend to zero in (4.55):

$$F_{ST}^{(0)} = \frac{P(1 - P)\{M + 4P(1 - P)\}}{P(1 - P)\{M + 4P(1 - P)\} + M(1 + M)}.$$  

Alternatively, this value could have been found from equation (4.8), where $ET$ and $ET_0$ are calculated analogously to equations (4.52) from results (3.59) to (3.61). We note that under weighting by size,

$$ET_0 = P ET_{11} + (1 - P) ET_{22} = c.$$  

Strobeck (1987) showed that under weighting by size, the mean coalescence time of two genes from the same subpopulation is, under quite general conditions, independent of the migration rates and pattern and equal to that of two genes from a panmictic population of the same total size (see also Theorem 5.1).

The value of the first order approximation $F_{ST}^{(1)}$ under weighting by size is calculated either from result (4.55) according to equation (4.9), or according to equation (4.10) where $E[T_3]$ and $E[T^2]$ are obtained by averaging results (3.62) to (3.64) analogously to (4.52):

$$F_{ST}^{(1)} = F_{ST}^{(0)} - P^2(1 - P)^2U \frac{8P(1 - P)(1 - 2P(1 - P) + M) + M^2}{[P(1 - P)\{M + 4P(1 - P)\} + M(1 + M)]^2}.$$  

Figure 4.21 shows that under weighting by size, $F_{ST}$ (given by equation (4.55)) depends much stronger on the relative sizes of the two colonies than under equal weighting (compare with figure 4.18). In particular, $F_{ST}$ under weighting by size (still maximal in the symmetric case of $P = 1/2$) converges to zero as an increasing proportion of the genes belong to the same subpopulation, while $F_{ST}$ calculated with equal weights on both colonies does not. This limiting behaviour under weighting by size is what one would expect intuitively, as in the limits of $P \downarrow 0$ and $P \uparrow 1$ the population approaches a panmictic population.

In figure 4.22, it is seen that (as was also the case under equal weighting) the decrease of $F_{ST}$ with increasing mutation rate $U$ under weighting by size is (when there is a fair amount of gene flow) much weaker under an asymmetric two-population model than when the colony sizes are equal. The dependence of $F_{ST}$ on $P$ is most important at small mutation rates and becomes weaker as the mutation rate increases. In contrast with the situation under equal weighting (compare with figure 4.20), $F_{ST}$ under weighting by size is approximately independent of the relative subpopulation sizes when the mutation rate is very large. In fact, it is seen from equation (4.55) that under weighting by size,

$$F_{ST} \sim \frac{1}{1 + U} \text{ as } \theta \to \infty$$  

(recall the notation "$\sim$" introduced on page 10), regardless of the relative subpopulation sizes.
Chapter 4: SUBPOPULATION DIFFERENTIATION

Figure 4.21: $F_{ST}$ values as a function of the relative subpopulation sizes in a general two-population model, under weighting by size.

The thick lines are for a migration rate of $M = 1$, the thin lines for $M = 5$. The solid lines are the values of the approximation $F_{ST}^{(0)}$. The long-dashed and dot-dashed lines are the exact $F_{ST}$ values assuming mutation rates of $U = 0.1$ and $U = 1$, respectively.

Figure 4.22: $F_{ST}$ and its approximations as a function of the scaled mutation rate, compared between a symmetric and an asymmetric two-population model under weighting by size.

The thin lines are the values of $F_{ST}$ and its approximations for a symmetric two-population model ($P = 1/2$), the thick lines those of a two-population model with $P = 1/20$. The exact $F_{ST}$ values are shown by a solid line, the approximations $F_{ST}^{(0)}$ and $F_{ST}^{(1)}$ by dotted and long-dashed lines, respectively. A migration rate of $M = 1$ was assumed.
4.3.2 The continental island model

We recall that the "migration rate" $M$ for this model is defined by equation (3.66), while the "mutation rate" $U$ is defined by equation (4.53). Under the neutral Wright-Fisher model for reproduction within the colonies (and in the coalescent approximation), $M$ is twice the number of genes exchanged between the continent and each of the islands every generation and $U$ is twice the number of mutations expected to occur in the total population every generation. All results are in terms of the parameters $M$, $U$, $c$, $n$ (the number of subpopulations, including the continent) and $P$, the proportion of the total population that lives in the continent.

The global value of $F_{ST}$ under equal weighting of the subpopulations follows from equation (4.3), where $f_0$ and $F$ are calculated from results (3.67) to (3.70) according to equations (4.51):

$$F_{ST} = \frac{\omega + P(n-1)^3 M^2 - 6Pn(n-1)^2MV}{\omega + P(n-1)^2 M \lambda + n(n-1)^2(1 - 4P^2)MV + n^2(n-1)^2(1 + 3P)M^2V + n^2(n-1)(3 + 2P)MV^2 + 2n(n-1)V^2 + 2n^2V^3} \quad (4.56)$$

where

$$V = P(1-P)U$$
$$\lambda = M(n^3 - n^2 - 1 + 2Pn + Pn^2 - 2P^2n^2) + n^2(n-1)M^2$$
$$\omega = P(n-1)^2 M \kappa + 2P(1-P)n(n-1)^2V + (n-1)^2(1 + 2P + 3Pn^2)MV + 2(n-1)(1 - 2Pn + Pn^2)V^2,$$

where in the latter equation

$$\kappa = (1-P)(2Pn^2 - 2 + n^2 - n - 4Pn + 4P).$$

Taking the limit of (4.56) as $\theta$ (and hence $U$) decreases to zero, we find the value of Slatkin's approximation $F_{ST}^{(0)}$ under equal weighting of the colonies:

$$F_{ST}^{(0)} = \frac{\kappa + M(n-1)^3}{\kappa + \lambda}. \quad (4.57)$$

This value could also have been found from equation (4.8), where $ET_0$ and $ET$ are calculated analogously to equations (4.51) from results (3.71) to (3.74). We note that, again, the mean coalescence time of two genes sampled from the same subpopulation, $ET_0$, under equal weighting of the subpopulations, depends on the migration rate. Inappropriately assuming in the calculation of $F_{ST}^{(0)}$ that $ET_0 = c$ (as is the case for the symmetric island and stepping-stone models) again leads to highly inaccurate $F_{ST}$ values.

The value of the first order approximation $F_{ST}^{(1)}$ under equal weighting is found from result (4.56) according to equation (4.9):

$$F_{ST}^{(1)} = F_{ST}^{(0)} - n(1-P)U \frac{\psi_0 + \psi_1 M + n(n-1)(n^2 - 2n + P)M^2}{(\kappa + \lambda)^2}.$$
Chapter 4: SUBPOPULATION DIFFERENTIATION

Figure 4.23: $F_{st}$ values as a function of the relative size of the continent in a continental island model, under equal weighting of the subpopulations.

The values of $F_{st}$ and $F_{st}^{(0)}$ are shown for a continental island model with $n = 3$ (figure a) and a continental island model with $n = 21$ (figure b). The level of gene flow between each island and the continent is the same in both figures and is also the same as in figure 4.18: the thick lines are for $M = 1$, while the thin lines correspond to $M = 5$. As in figure 4.18, the solid lines indicate the approximation $F_{st}^{(0)}$, while the long-dashed lines show the exact $F_{st}$ values for $U = 0.1$ and the dot-dashed lines those for $U = 1$.

where $\kappa$ and $\lambda$ are as above and where in addition

$$
\psi_0 = (1 - P)(n^2 - n - 2 - 4Pn - 24P^2n + 16P^3n + 4Pn^2 + 6P^2n^2 - 4P^3n^2 - 8P + 32P^2 - 16P^3)
$$

$$
\psi_1 = (1 - P)(2n^3 + 4Pn^3 - 4n^2 - 5Pn^2 - n - 6Pn + 8P^2n + 2).
$$

This result could also have been obtained from equation (4.10), calculating $E[T_0^2]$ and $E[T^2]$ analogously to (4.51) from results (3.75) to (3.78).

Figure 4.23 illustrates the dependence of $F_{st}$ and $F_{st}^{(0)}$ (under equal weighting) on the relative sizes of continent and islands, for a continental island model consisting of a continent and two islands (figure 4.23a) and one with twenty islands (figure 4.23b). As was the case for the two-population model (see figure 4.18), which is in fact a continental island model with only one island, this dependence is seen to be weak, for the range of mutation rates shown, particularly when there are many islands or a high level of gene flow.

Figure 4.24 shows the $F_{st}$ values under equal weighting in a continental island model, plotted against the number of subpopulations, for different relative sizes of the continent. For interest, we have also plotted the corresponding $F_{st}$ values under the symmetric island model described in Section 3.1 (indicated by pluses). Both the “mutation rate” $U = c/2$ (where $c = n$ in the case of the symmetric island model) and the “migration rate” $M$ are assumed to take constant values (independent of the number of islands) which are the same under both models. In the case of the
Figure 4.24: $F_{ST}$ as a function of the number of subpopulations in the continental island model, compared between different relative sizes of the continent and compared with the symmetric island model (equal weighting). The filled symbols, joined by a line, represent the exact $F_{ST}$ value of a continental island model consisting of a continent and $n-1$ islands, for $P = 1/100$ (diamonds), $P = 1/2$ (circles) and $P = 99/100$ (squares). The pluses are the exact $F_{ST}$ values of a symmetric island model with $n$ colonies. All models assume a migration rate of $M = 1$: every two generations, there is, under the Wright-Fisher model, one gene exchanged between each island and the continent (in the case of a continental island model) or between each island and the rest of the population (in the case of the symmetric island model). In all cases a mutation rate of $U = c\theta = 0.1$ was assumed (where $c = n$ for the symmetric island model): under the Wright-Fisher model, one mutation is expected in the total population every 20 generations.
neutral Wright-Fisher model for reproduction within the colonies, \( U \) is both under the symmetric and the continental island models twice the expected number of mutations in the total population every generation; for the continental island model, \( M \) (defined by equation (3.66)) is twice the number of genes exchanged between each island and the continent, every generation, while for the symmetric island model, \( M \) (defined in Section 3.1) is twice the number of genes exchanged between each island and the rest of the population, every generation. Surprisingly, the \( F_{ST} \) value of the symmetric island model is (for small mutation rates) a reasonable approximation for that of the continental island model, particularly when there are many islands. Indeed, as the number of colonies becomes large, \( F_{ST}^{(0)} \) approaches the same limiting value under both models, namely Wright’s value of \( 1/(1+M) \), independent of the relative sizes of continent and islands (this follows from results (4.57) and (4.16)). If \( U \) is held constant as \( n \to \infty \), \( F_{ST} \) and \( F_{ST}^{(1)} \) also converge to this value, under both models. We have checked that, for small mutation rates, the \( F_{ST} \) value under the symmetric island model is generally a good approximation for the corresponding value under the continental island model, except when there are only a few islands and the continent is much larger than the islands and is isolated (\( M \) less than 1, say). For \( n = 2 \) and \( P = 99/100 \), both values differ by as much as a factor of two for \( M \approx 0.1 \) and small mutation rate (see figure 4.19).

Figure 4.25 illustrates the dependence of \( F_{ST} \) on the mutation rate \( U \) and the performance of the approximations \( F_{ST}^{(0)} \) and \( F_{ST}^{(1)} \), under a continental island model with five islands and equal weighting of the subpopulations. The extent to which \( F_{ST} \) depends on the mutation rate varies considerably with the relative size of the continent. While the decrease of \( F_{ST} \) with increasing mutation rate is never rapid under a continental island model, it is much slower as the relative size of the continent is larger. As a consequence, both approximations \( F_{ST}^{(0)} \) and \( F_{ST}^{(1)} \) are much more accurate when the continent is large than in the case of a small continent. Whereas for small mutation rates, the dependence of \( F_{ST} \) on \( P \) is weak and gets even weaker as there are more colonies (see also figure 4.23), it is seen that on the contrary, \( F_{ST} \) under equal weighting depends very strongly on the relative sizes of continent and islands when the mutation rate is very large. The exact \( F_{ST} \) values of the two models shown differ by up to a factor of 14 at very large mutation rates. This difference can be larger if the number of colonies is larger.

In the previous section it was observed that (for the symmetric models studied there) \( F_{ST} \) depends more strongly on the mutation rate as gene dispersal is more restricted. For example, the decrease of \( F_{ST} \) with increasing mutation rate was seen to be faster under a stepping-stone model than under a symmetric island model. Comparing the continental island model with the symmetric island model, we have checked that, depending on the relative size of the continent, the decrease of \( F_{ST} \) with increasing mutation rate under the continental island model can be both faster or slower than under the corresponding symmetric island model. Under a continental island model with \( n = 3 \) and \( P = 1/3 \) (which is a linear stepping-stone model with three subpopulations),
Figure 4.25: $F_{ST}$ and its approximations as a function of the scaled mutation rate in the continental island model, compared between different sizes of the continent (equal weighting).

The values of $F_{ST}$ and its approximations are shown for a continental island model with $n = 6$ and $M = 1$. The thin lines correspond to $P = 1/6$ (equal-sized colonies), the thick lines to $P = 19/20$ (large continent). The solid lines are the exact $F_{ST}$ values, while the dotted lines indicate Slatkin's approximation $F_{ST}^{(6)}$ and the long-dashed lines represent the first order approximation $F_{ST}^{(1)}$. 
Chapter 4: SUBPOPULATION DIFFERENTIATION

$F_{ST}$ decreases more rapidly with increasing mutation rate than under the symmetric island model (which is for $n = 3$ equivalent to the circular stepping-stone model), as one would have expected intuitively since under the linear stepping-stone model gene dispersal is more restricted than under the circular one. As the relative size of the middle colony decreases, the decrease of $F_{ST}$ with increasing mutation rate becomes even faster than under the linear stepping-stone model with $n = 3$.

Because (under equal weighting) the $F_{ST}$ values of the continental and symmetric island models are close to each other when the mutation rate is small, and because the rate of decrease of $F_{ST}$ with increasing mutation rate can be very different under the two models (depending on the relative size of the continent in the continental island model), the $F_{ST}$ value of the symmetric island model is often very different from (and cannot serve anymore as an approximation for) that under the continental island model when the mutation rate is very large.

We also calculate the global values of $F_{ST}$ and its approximations under weighting by size. Calculating $f_0$ and $f$ according to equations (4.52), from results (3.67) to (3.70), and substituting the resulting values into equation (4.3), we find the value of $F_{ST}$ under weighting by size:

$$F_{ST} = \frac{R + 2P^2(1 - P)^3(n - 1)U^2 + P(1 - P)^2(3n^2 + 2P - 5)WU + (n - Pn + 3P - 2)W^2}{R + X_0 + X_1 W + (n - 3P^2U + 2PU + U + P - 1)W^2 + W^3}$$

(4.58)

where

$$W = (n - 1)M$$
$$R = 2P(1 - P)^2(n + Pn - 2)U + (1 - P)^2(3Pn - 6P + n - 2 + 4P^2)W - P^2W^2$$
$$X_0 = 2P(1 - P)^2 \{P(1 - P)U + P^2 + Pn - 3P + 1\} U^2$$
$$X_1 = (1 - P)(3PU - P^2U - 2P^2U + 3Pn - 3P - 3P^2 + 2P^3 + 1)U.$$

The value of $F_{ST}^{(0)}$ under weighting by size is obtained either from result (4.58) according to equation (4.7), or by averaging the mean coalescence times (3.71) to (3.74) analogously to equations (4.52) and substituting the resulting values of $ET_0$ and $ET$ into equation (4.8):

$$F_{ST}^{(0)} = \frac{(1 - P)^2(3Pn - 6P + n - 2 + 4P^2) + (1 - P)(P + n - 2)W}{(1 - P)^2(3Pn - 6P + n - 2 + 4P^2) + P(1 - P)W + W(W + n - 1)}.$$

We note that under weighting by size, the mean coalescence time of two genes sampled from the same subpopulation satisfies

$$ET_0 = PE_T^{(0)} + (1 - P)ET_{11} = c.$$

The value of the first order approximation $F_{ST}^{(1)}$ under weighting by size is found from result (4.58) according to equation (4.9). Alternatively it is calculated according to equation (4.10), where $E[T_0^2]$ and $E[T^2]$ are obtained from results (3.75) to (3.78) analogously to equations (4.52).
Figure 4.26: $F_{ST}$ values as a function of the relative size of the continent in a continental island model under weighting by size.

The $F_{ST}$ values under weighting by size are shown for a continental island model with $n = 3$ (figure a) and a continental island model with $n = 21$ (figure b). The migration and mutation rates are as in figure 4.23.

The result is:

$$F_{ST}^{(1)} = F_{ST}^{(0)} - (1 - P)^2 U \frac{Z_0 + Z_1 W + (n + P^2 - 2)W^2}{\{(1 - P)^2(3Pn - 6P + n - 2 + 4P^2) + P(1 - P)W + W(W + n - 1)\}^2}$$

where $W$ is as above and where in addition

$$Z_0 = (1 - P)(n - 1)(24P^4 + 7P^3n - 33P^3 + P^2n - 3P^2 + 3P + 1) + (1 - P)^3(16P^2 - 8P^2 - 5P - 1)$$

$$Z_1 = 2(1 - P)(4P^3 + P^2n - 2P^2 + 2Pn - 4P + n - 2).$$

As observed for the two-population model and illustrated in figure 4.26, the dependence of $F_{ST}$ on the relative colony sizes in the continental island model is, for small mutation rates, much stronger under weighting by size than under equal weighting. In particular, the $F_{ST}$ values under weighting by size converge to zero (which corresponds to a panmictic population) as an increasing proportion of the genes belong to the continent ($P \uparrow 1$). The coefficient $F_{ST}$ decreases more slowly with increasing mutation rate as the relative size of the continent is larger (except possibly when the colonies are nearly isolated). When the mutation rate is very large (no values shown), $F_{ST}$ under weighting by size is approximately independent of the relative sizes of continent and islands, and is also approximately the same as the $F_{ST}$ value under the corresponding symmetric island model. This is opposed to the situation under equal weighting. For small mutation rates, the $F_{ST}$ value of the continental island model under weighting by size is very different from that of the symmetric island model if the continent is much larger than the islands.
4.3.3 Four-population model with geographic barrier

We recall that according to Subsection 3.3.3, \( M \) is the total scaled migration rate from each colony, where time-scaling is such that two genes in the same colony coalesce at rate 1. Each gene leaves its colony at rate \( M/2 \) and mutates at rate \( \theta/2 \), \( \theta \) being the scaled mutation rate. In the case of the neutral Wright-Fisher model for reproduction within the subpopulations, \( M = \lim_{N \to \infty} (4Nm) \) and \( \theta = \lim_{N \to \infty} (4N\mu) \), where \( 2N \) is the number of genes per subpopulation, \( m \) is the proportion of each subpopulation that is replaced by immigrants every generation and \( \mu \) is the probability of mutation per gene per generation. In that case, \( M \) and \( \theta \) are, in the coalescent approximation, twice the number of migrant genes and twice the expected number of mutant genes, respectively, per subpopulation per generation. Recall also that \( \epsilon \) is the proportion of the immigrants into each subpopulation that come from across the "barrier" which separates colonies 1 and 2 from colonies 3 and 4 \((0 < \epsilon < 1)\).

As all colonies have the same size in this model, equal weighting is equivalent to weighting by size. The global value of \( F_{ST} \) is calculated according to equation (4.3), where \( f_0 \) is given by result (3.79), with \( s = \theta \), and where

\[
\tilde{f} = \frac{1}{4} \left( E[e^{-\theta T_{11}}] + E[e^{-\theta T_{13}}] + E[e^{-\theta T_{14}}] + E[e^{-\theta T_{12}}] \right)
\]

(4.59)
is calculated from results (3.79) to (3.82). We find:

\[
F_{ST} = \frac{4M^2 \{1 + \epsilon(1 - \epsilon)\} + 8M\theta + 3\theta^2}{4M^2 \{1 + \epsilon(1 - \epsilon)\} + 8M\theta + 3\theta^2 + 16\epsilon(1 - \epsilon)M^2(2M + \theta) + 16M\theta(M + \theta) + 4\theta^3}.
\]

(4.60)
The values of Slatkin's approximation \( F_{ST}^{(0)} \) and the first order approximation \( F_{ST}^{(1)} \) are found from equations (4.7) and (4.9), letting \( \theta \perp 0 \) in result (4.60) and in its derivative with respect to \( \theta \):

\[
F_{ST}^{(0)} = \frac{1 + \epsilon(1 - \epsilon)}{1 + \epsilon(1 - \epsilon) + 8\epsilon(1 - \epsilon)M}
\]

\[
F_{ST}^{(1)} = F_{ST}^{(0)} - 4\theta \frac{(1 - \epsilon)(1 - \epsilon)^2}{(1 + \epsilon(1 - \epsilon) + 8\epsilon(1 - \epsilon)M)^2}.
\]

Alternatively these values could have been obtained from equations (4.8) and (4.10), where \( ET_0 \) and \( E[T_2] \) are given by equations (3.83) and (3.87), respectively, and where \( ET \) and \( E[T^2] \) are calculated from results (3.83) to (3.90) analogously to equation (4.59).

In figure 4.27, we plot the values of \( F_{ST} \) and its approximation \( F_{ST}^{(0)} \) against \( \epsilon \), which indicates the level of asymmetry in the population structure, for two different levels of gene flow and for different values of the mutation rate. It is observed that the amount of genetic differentiation among the subpopulations, as measured by \( F_{ST} \), is minimal in the symmetric case without barrier (i.e. the circular stepping-stone model; \( \epsilon = 1/2 \)) and increases as the level of asymmetry induced by the barrier increases. This is because, as \( \epsilon \) deviates further from 1/2, two genes from the same colony become more likely to be identical by descent, while two genes sampled at random from
Figure 4.27: $F_{ST}$ values under the four-population model with geographic barrier, as a function of the fraction of the gene flow that is across the barrier.

The thick lines are for $M = 1$, the thin lines for $M = 10$. The solid lines are the values of the approximation $F_{ST}^{(0)}$. The long-dashed lines are the exact $F_{ST}$ values for $\theta = 0.1$, the dot-dashed lines are those for $\theta = 0.5$.

Note also that the approximation $F_{ST}^{(0)}$ becomes considerably less accurate as the barrier is more important.

Figure 4.28 shows the decrease of $F_{ST}$ with increasing level of gene flow for different values of $\epsilon$. The solid, long-dashed and dot-dashed lines assume respectively that half, 10% or 1% of the total amount of gene flow is across the barrier. The limiting value of $F_{ST}$ as $M$ decreases to zero is independent of the value of $\epsilon$, which is expected intuitively since this limiting situation corresponds to the complete isolation of all four subpopulations. An increasing (total) level of gene flow has a stronger homogenizing effect on the genetic composition of the population as the model is more symmetric. If the fraction of the gene flow that goes across the barrier is very small, the (global) $F_{ST}$ value can still be high even when there is a lot of gene flow on the whole. If one were not aware of the presence of a substantial barrier and calculated $F_{ST}$ values assuming $\epsilon = 1/2$ (while the actual structure of the population is highly asymmetric), a considerable error may be made: for large $M$ and small $\theta$, the $F_{ST}$ value for $\epsilon = 1/10$ is more than twice that of the symmetric
Figure 4.28: $F_{ST}$ as a function of the scaled migration rate under the four-population model with geographic barrier, for different levels of asymmetry.

The values shown are the exact $F_{ST}$ values at a scaled mutation rate of $\theta = 0.01$ and scaled migration rate $M$. The solid line is for $\epsilon = 1/2$ (the circular stepping-stone model with $n = 4$), the long-dashed line assumes $\epsilon = 1/10$ and for the dot-dashed line it was assumed that $\epsilon = 1/100$.

model, while that for $\epsilon = 1/100$ is up to 20 times larger than the $F_{ST}$ value of the symmetric model.

In figure 4.29 the decrease of $F_{ST}$ with increasing mutation rate is compared between a symmetric (thin lines) and a highly asymmetric (thick lines) four-population model. Whereas the dependence of $F_{ST}$ on the mutation rate is already fairly important under the symmetric model, which is the circular stepping-stone model with four colonies (see also the previous section), $F_{ST}$ decreases even faster with increasing mutation rate when there is a barrier. As a consequence, both approximations $F_{ST}^{(0)}$ and $F_{ST}^{(1)}$, which are already not too accurate under the circular stepping-stone model with $n = 4$, are much less accurate under the asymmetric model. For $\theta = 0.25$ under the asymmetric model (that is under the Wright-Fisher model, one mutation expected in the total population every two generations), Slatkin's approximation $F_{ST}^{(0)}$ is more than 50% larger than the exact $F_{ST}$ value, while the first order approximation $F_{ST}^{(1)}$ is 40% smaller than $F_{ST}$. While the dependence of $F_{ST}$ on $\epsilon$ is very important for small mutation rates, the value of $F_{ST}$ is, on the contrary, approximately independent of $\epsilon$ (and approximately the same as the $F_{ST}$ values of both the circular stepping-stone and symmetric island models with four colonies) when the mutation rate is very large.
Figure 4.29: $F_{ST}$ and its approximations as a function of the scaled mutation rate, compared between four-population models with and without geographic barrier.  

The thick lines are the exact and approximate $F_{ST}$ values under the four-population model with $\epsilon = 1/20$, the thin lines are those with $\epsilon = 1/2$. In both cases, the exact $F_{ST}$ value is indicated by a solid line, the dotted line is the approximation $F_{ST}^{(0)}$, and the first order approximation $F_{ST}^{(1)}$ is plotted by a long-dashed line. A scaled migration rate of $M = 1$ is assumed.
4.3.4 Discussion

We have illustrated how asymmetries in the population structure can affect the amount of subpopulation differentiation, as measured by $F_{ST}$, in very different ways. In the two-population and continental island models, unequal colony sizes have at small mutation rates only a weak effect on the value of $F_{ST}$, calculated with equal weight on all subpopulations, except possibly when there are only a few colonies which are at the same time fairly isolated. Unequal colony sizes do strongly affect the $F_{ST}$ value under equal weighting when the mutation rate is very large (under the Wright-Fisher model: more than one mutation in the total population per generation). If all subpopulations are weighted by their relative sizes, the effect of unequal subpopulation sizes is, on the contrary, important for small mutation rates but weak when the mutation rate $U$ is much larger than both the migration rate ($M$) and 1. Under both weightings, unequal colony sizes have a strong impact on the rate with which $F_{ST}$ decreases with increasing mutation rate. In the two-population model under equal weighting, unequal subpopulation sizes weaken the dependence of $F_{ST}$ on the mutation rate, increasing the accuracy of both approximations $F_{ST}^{(0)}$ and $F_{ST}^{(1)}$. In the continental island model under equal weighting with more than two colonies, the decrease of $F_{ST}$ with increasing mutation rate is slower (and hence the approximations $F_{ST}^{(0)}$ and $F_{ST}^{(1)}$ are more accurate) as a larger proportion of the population belongs to the continent. Unequal colony sizes can both decrease or increase the value of $F_{ST}$. If in the continental island model, all subpopulations are equally-sized (so that it is only the migration pattern which makes the continent different from the islands), the $F_{ST}$ value of the continental island model is close to that of the symmetric island model, at all mutation rates. Of course, other characteristics are likely to be very different between the two models. Studying an example of a population structure with unequal migration rates, we have shown that unequal migration rates can substantially increase both the (global) $F_{ST}$ value and the dependence of $F_{ST}$ on the mutation rate as particular migration rates become relatively small and groups of subpopulations become increasingly isolated from each other. In such cases, Slatkin's approximation $F_{ST}^{(0)}$ and the first order approximation $F_{ST}^{(1)}$ can be highly inaccurate at realistic mutation rates.

In the previous section we observed that, for the symmetric models of population structure considered there (which all assumed equal-sized subpopulations), $F_{ST}$ depends strongly on the precise structure of the population at small mutation rates, but is approximately independent of the details of population structure when the mutation rate is very large. At present we have illustrated that this is not true for models with unequal-sized colonies, if all colonies are assigned equal weight. However, the above property does hold in the case of unequal colony sizes if all subpopulations are weighted by their relative sizes. We will prove in Chapter 5 that when the mutation rate becomes very large, the global value of $F_{ST}$ under weighting by size is approximately independent of the details of population structure and behaves asymptotically as $1/(1 + U/(n - 1))$ (see Corollary 5.3
for a precise formulation of this result), where \( U = c\theta \) is, under the Wright-Fisher model, twice the number of mutations expected in the total population per generation (for the models with equal-sized subpopulations, \( c = n \) and \( U = n\theta \)). This fact may be of interest when studying loci with very high mutation rates in populations where the different subpopulations are very isolated. The rate at which \( F_{ST} \) decreases with increasing mutation rate at small mutation rates always depends strongly on the precise structure of the population.

We observed in the previous section that, among the symmetric models studied there, the dependence of \( F_{ST} \) on the mutation rate is (for a fixed number of subpopulations) maximal under the one-dimensional stepping-stone model, where gene dispersal is maximally restricted. In this section, we have illustrated that unequal migration rates or subpopulation sizes under restricted dispersal can cause the decrease of \( F_{ST} \) with increasing mutation rate to be even more rapid than under the one-dimensional stepping-stone model. Under the four-population model with barrier \( (\epsilon^1/2) \), the absolute value of the derivative of \( F_{ST} \) with respect to the scaled mutation rate \( \theta \) at \( \theta = 0 \) is larger (much larger, if the model is highly asymmetric) than under the circular stepping-stone model with four colonies. For three colonies in a line, with the middle colony smaller than the outer ones (this is a continental island model with \( n = 3 \) and \( P < 1/3 \)), the absolute value of the derivative of \( F_{ST} \) with respect to the mutation rate at mutation rate zero is (both under equal weighting and under weighting by size) larger than under the linear stepping-stone model with three colonies (which is a continental island model with \( n = 3 \) and \( P = 1/3 \)), where it is in turn larger than under the circular stepping-stone model with three colonies. The bounds on the discrepancy between \( F_{ST}^{(0)} \) and \( F_{ST} \) derived in the previous section on the basis of the circular stepping-stone model (inequalities (4.47) and (4.48)), do not hold for the four-population model with barrier. For three colonies in a line, these error bounds (with \( n\theta \) replaced by \( U = c\theta \)) are valid if \( M \) is defined as under the continental island model (i.e. the scaled migration rate between neighbouring colonies is \( M \)), but not if the scaled migration rate between adjacent colonies is only \( M/2 \) (which is the usual setting of a stepping-stone model). However, the bounds (4.49) and (4.50) on the absolute and relative errors made by using \( F_{ST}^{(0)} \) as an approximation for \( F_{ST} \) are still valid for the less symmetrical models studied in this section, if all colonies are given equal weight (this can be shown from the analytical values calculated in this section):
the same model of population structure) is subject to:

\[ |F_{ST}(U_1) - F_{ST}(U_2)| \leq |U_1 - U_2| \left( F_{ST}^{(1)} \right)^2. \]  

(4.62)

However, it appears hard to prove the inequalities (4.61) and (4.62) under an arbitrary model of population structure and it is not known whether they hold generally under equal weighting. They are not valid in general under weighting by size.

For all the models of population structure studied in this chapter, \( F_{ST} \) is (both under equal weighting and under weighting by size) a monotonically decreasing and convex function of the mutation rate, so that the approximations \( F_{ST}^{(0)} \) and \( F_{ST}^{(1)} \) are, respectively, an upper and a lower bound for \( F_{ST} \). Again, it is not known whether this fact is true in general.

Summarizing our observations, we conclude that asymmetries in the population structure can have an important impact on the \( F_{ST} \) value when groups of subpopulations are relatively isolated from each other, when the overall level of gene flow is low, or, if \( F_{ST} \) is calculated with equal weights on all subpopulations, at very high mutation rates. Under weighting by size, asymmetries can strongly affect the value of \( F_{ST} \) at small mutation rates, while little effect on \( F_{ST} \) is seen at very high mutation rates. We have illustrated that (under both weightings) asymmetries have a strong effect on the rate with which \( F_{ST} \) decreases with increasing mutation rate. Slatkin's approximation \( F_{ST}^{(0)} \) seems accurate in most realistic situations, but can be inaccurate at highly mutable loci or when some subpopulations are relatively isolated from each other, either by the presence of some kind of barrier or because dispersal is restricted to an essentially one-dimensional or very large two-dimensional habitat (see also the previous section).
Chapter 5

GENERAL RESULTS

In Chapters 3 and 4, coalescence times and $F_{ST}$ values were calculated for specific models of population structure. Some results were highlighted which are the same for various population structures. In this chapter, several of the observations made are proved with more generality. Section 1 is concerned with some basic identities on coalescence times and probabilities of identity by descent, valid under weighting by size. Section 2 deals with the asymptotic behaviour of $F_{ST}$ as the mutation rate becomes large. In Section 3 some graph-theoretic conditions on the population structure are set out under which the mean coalescence time of a pair of individuals from the same subpopulation is independent of the migration rate. In Section 4 it is shown that for a range of sufficiently symmetric models of population structure there is a uniform relationship between the migration rate and the value of Slatkin's approximation $F_{ST}^{(0)}$ calculated for a pair of neighbouring subpopulations.

Throughout this chapter it is assumed that the genealogy of a sample from the population is well described by the structured coalescent, given by equation (2.2), where the $c_i$ correspond to the relative subpopulation sizes, where

$$\forall i \in S : M_i = \sum_{j \neq i} M_{ij} < \infty$$

and where migration is assumed to be "conservative", that is,

Assumption 5.1

$$\forall i \in S : \sum_{j \neq i} c_i M_{ij} = \sum_{i \neq j} c_j M_{ji}$$

(cf. Assumption 2.1). Assumption 5.1 is satisfied if the subpopulation sizes are not affected by migration. When concentrating on genealogy, we talk about "individuals" rather than genes. Recall however that these individuals are assumed to be haploid, for example genes (see Chapter 2).

We denote by $T_{ij}$ the coalescence time of an individual from subpopulation $i$ and an individual from subpopulation $j$, and by $f_{ij}$ the probability of identity by descent of a gene from subpopulation $i$ and one from subpopulation $j$. We recall that $c := \sum_{i \in S} c_i$ and we use the notation...
Chapter 5: GENERAL RESULTS

\[ U = ce, \text{ which is, under the Wright-Fisher model for reproduction within subpopulations, twice} \]
\[ \text{the number of mutations expected in the total population per generation. As was also the case in} \]
\[ \text{the previous chapter, the results on } F_{ST} \text{ obtained in this chapter are exact only for the infinite} \]
\[ \text{alleles model for mutation, in which the probability of identity of a pair of genes is equal to their} \]
\[ \text{probability of identity by descent.} \]

5.1 Genealogical results under weighting by size.

In Chapter 3 it was observed that the mean coalescence time of two individuals from the same subpopulation is often independent of the migration rate and, in the time-scaling used, equal to \( c \). Strobeck (1987) proved that the mean coalescence time of two individuals from the same subpopulation is equal to \( c \) under any model of population structure (satisfying Assumption 5.1) if the various subpopulations are weighted by their relative sizes. The following theorem states this result (identity (ii)), together with related results on the probability of identity by descent of a pair of genes and the moments of higher order of the coalescence time of a pair of individuals. As defined in the previous chapter, \( f_0 \) denotes the probability of identity by descent of a pair of genes sampled at random from a single subpopulation, while \( f \) is that of two genes sampled from the total population. Similarly, \( T_0 \) and \( T \) denote the coalescence times of, respectively, two individuals from a single subpopulation and two individuals from the total population.

**Theorem 5.1** Assume the population consists of a finite number of subpopulations. Weighting each subpopulation by its relative size (see equations \((4.52)\)), the following identities hold:

(i) \[ f_0 + U f = 1. \]

(ii) If \( \forall i, j \in S: ET_{ij} < \infty \), then \[ ET_0 = c \]

(Strobeck 1987).

(iii) If \( \nu \in N_0 \) and \( \forall i, j \in S: E[T_{ij}^\nu] < \infty \), then \[ E[T_0^\nu] = \nu c E[T^{\nu-1}]. \]

We proved in Chapter 2 that the condition of identity (ii), that every two individuals in the population have finite mean coalescence time, is fulfilled if the Markov chain generated by the scaled migration rates is irreducible (Theorem 2.3). An equivalent condition will be given in Section 3.

**Proof of the theorem:** As the probability of identity by descent of a pair of genes is the moment-generating function of their coalescence time, evaluated in the scaled mutation rate \( \theta \) (see
Chapter 5: GENERAL RESULTS

the argument preceding equation (4.4)), equations (2.44) can for a sample of two individuals and for \( s = \theta \) be written as

\[
\left( \frac{\delta_{ij}}{P_i c} + \frac{M_i}{2} + \frac{M_j}{2} + \theta \right) f_{ij} - \sum_{k \in S: k \neq i} \frac{M_{ik}}{2} f_{kj} - \sum_{k \in S: k \neq j} \frac{M_{jk}}{2} f_{ik} = \frac{\delta_{ij}}{P_i c} \tag{5.2}
\]

\((i, j \in S)\), where \( P_i = c_i / c \) is the proportion of the population that lives in subpopulation \( i \).

Multiplying equation (5.2) by \( P_i P_j \) and summing the resulting equation over all \( i, j \in S \), we obtain:

\[
\sum_{i \in S} \sum_{j \in S} P_i P_j \left\{ \left( \frac{\delta_{ij}}{P_i c} + \frac{M_i}{2} + \frac{M_j}{2} + \theta \right) f_{ij} - \sum_{k \neq i} \frac{M_{ik}}{2} f_{kj} - \sum_{k \neq j} \frac{M_{jk}}{2} f_{ik} \right\} = \sum_{i \in S} \sum_{j \in S} P_i P_j \delta_{ij} \frac{P_i}{P_i c},
\]

which simplifies to

\[
\frac{1}{c} \sum_{i \in S} P_i f_{ii} + \sum_{i \in S} \sum_{j \in S} P_i P_j M_{ij} f_{ij} + \theta \sum_{i \in S} \sum_{j \in S} P_i P_j f_{ij} - \sum_{i \in S} \sum_{j \in S} P_i P_j \sum_{k \neq i} M_{ik} f_{kj} = \frac{1}{c}. \tag{5.3}
\]

Changing the order of the summations over \( i \) and \( k \) in the fourth term of the left-hand side and using Assumption 5.1, this term can be written as

\[
\sum_{i \in S} \sum_{j \in S} P_i P_j \sum_{k \neq i} M_{ik} f_{kj} = \sum_{i \in S} \sum_{j \in S} P_j f_{kj} \sum_{i \neq k} P_i M_{ik} = \sum_{i \in S} \sum_{j \in S} P_j f_{kj} P_k M_k.
\]

Hence in the left-hand side of equation (5.3), the fourth term cancels against the second term. Multiplied by \( c \), the remaining equation is

\[
\sum_{i \in S} P_i f_{ii} + c \theta \sum_{i \in S} \sum_{j \in S} P_i P_j f_{ij} = 1.
\]

Substituting the definitions of \( f_0 \) and \( f \) under weighting by size, (4.52), into this equation, identity (i) is obtained. The identities (ii) and (iii) follow by differentiating result (i) with respect to \( \theta \) once and \( \nu \) times, respectively, and taking the limit as \( \theta \to 0 \).

Theorem 5.1 lists some useful relationships between samples from a single subpopulation and samples from the total population, valid under weighting by size. These relationships suggest that \( F_{ST} \) under weighting by size may be estimated from a random sample of genes from the total population (all individuals in the population being equally likely to be sampled; no knowledge of which individuals belong to which subpopulations is needed) and an estimate of the number of mutations expected in the total population every generation. Alternatively, if it is reasonable to assume that all subpopulations are identical with respect to size, migration and mutation, a sample of genes from one subpopulation, together with an estimate of the expected number of mutations in the total population per generation, may be sufficient to estimate \( F_{SR} \). If reliable estimates of both \( f_0 \) and \( f \) under weighting by size are available, these yield according to result (i)
Chapter 5: GENERAL RESULTS

an estimate of the mutation rate $U = c \theta$. We stress that the identities in Theorem 5.1 do not hold under weightings other than by size or when not all subpopulations are sampled (as is the case for pairwise $F_{ST}$ values or when there is gene flow from a subpopulation which has not been considered as belonging to the population). For example, the mean coalescence time of a pair of individuals sampled from the same subpopulation under equal weighting in the two-population and continental island models depends on the migration rate (see Chapter 3). Note also that Theorem 5.1 heavily relies on Assumption 5.1 that the subpopulation sizes are maintained under migration.

5.2 The asymptotic behaviour of $F_{ST}$ at large mutation rates.

In Chapter 4 it was observed that at very large mutation rates, $F_{ST}$ under weighting by size is approximately independent of the details of population structure (migration pattern, migration rates and relative subpopulation sizes) and dependent only on the number of mutations expected in the total population every generation and on the number of subpopulations. This asymptotic behaviour of $F_{ST}$ under weighting by size as the mutation rate becomes large is established in the following theorem in a more general setting. Recall that $S$ denotes the set of the subpopulation labels. The theorem is concerned with the $F_{ST}$ value, $F_{ST}(C)$, of a collection of subpopulations with labels in $C \subseteq S$. Weighting the subpopulations with labels in $C$ by their relative sizes,

$$F_{ST}(C) = \frac{f_0(C) - \bar{f}(C)}{1 - \bar{f}(C)}$$

where, denoting $c(C) := \sum_{i \in C} c_i$,

$$\begin{align*}
  f_0(C) &= \sum_{i \in C} \frac{c_i}{c(C)} f_{ii} \\
  \bar{f}(C) &= \sum_{i \in C} \sum_{j \in C} \frac{c_i}{c(C)} \frac{c_j}{c(C)} f_{ij}.
\end{align*}$$

(5.4)

We use the notation $\text{card}(C)$ for the number of subpopulations with labels in $C$ and we denote

$$U(C) := c(C) \theta.$$  

(5.5)

If $\theta = 4N\mu$, as is the case under the Wright-Fisher model for reproduction within subpopulations, $U(C)$ is twice the number of mutations expected per generation in the collection of subpopulations with labels in $C$. Recall the notation "$\sim$" introduced on page 10.
Chapter 5: GENERAL RESULTS

Theorem 5.2 If $C \subseteq S$ is finite and the subpopulations in $C$ are weighted by their relative sizes, then

$$F_{ST}(C) \sim \frac{1}{1 + U(C)/\{card(C) - 1\}} \text{ as } \theta \to \infty.$$ 

Corollary 5.3

(i) If the population consists of $n$ subpopulations $(n < \infty)$, then the global value of $F_{ST}$ under weighting by size satisfies

$$F_{ST} \sim \frac{1}{1 + U/(n - 1)} \text{ as } \theta \to \infty.$$ 

(ii) The $F_{ST}$ value of a pair of subpopulations with labels $i$ and $j \in S$ $(i \neq j)$ with $c_i = c_j = 1$ satisfies

$$F_{ST}(i,j) \sim \frac{1}{1 + 2\theta} \text{ as } \theta \to \infty.$$ 

Proof of the corollary: Statements (i) and (ii) follow from Theorem 5.2 by choosing $C = S$ and $C = \{i,j\}$, respectively. □

Proof of Theorem 5.2: We let $\theta$ tend to infinity in equation (5.2). Because by the dominated convergence theorem, $\lim_{\theta \to \infty} f_{ij} = \lim_{\theta \to \infty} E[e^{-\theta T_{ij}}] = 0$ for every $i,j \in S$ and because $\sum_{k \in S \setminus \{i\}} M_{ik} = M_i < \infty$ for every $i \in S$ (equation (5.1)), we obtain (again by the dominated convergence theorem if $S$ is infinite) that

$$\lim_{\theta \to \infty} (\theta f_{ij}) = \frac{\delta_{ij}}{c_i}$$

for all $i,j \in S$. Hence, using equations (5.4) and (5.5),

$$\lim_{\theta \to \infty} \{U(C)f_0(C)\} = \lim_{\theta \to \infty} \left\{ \sum_{i \in C} c_i \theta f_{ii} \right\} = card(C)$$

and

$$\lim_{\theta \to \infty} \{U(C)f(C)\} = \lim_{\theta \to \infty} \left\{ \sum_{i \in C} \sum_{j \in C} c_i c_j \frac{\theta f_{ij}}{c(C)} \right\} = 1.$$

Since $\lim_{\theta \to \infty} f_0(C) = \lim_{\theta \to \infty} f(C) = 0$, it follows that

$$\lim_{\theta \to \infty} \frac{F_{ST}(C)}{1 + U(C)/\{card(C) - 1\}} = \lim_{\theta \to \infty} \left\{ \frac{card(C) - 1 + U(C)}{card(C) - 1} \cdot \frac{f_0(C) - f(C)}{1 - f(C)} \right\}$$

$$= \lim_{\theta \to \infty} \frac{U(C)f_0(C) - U(C)f(C)}{card(C) - 1}$$

$$= 1.$$ 

□

The intuition behind this result (and behind the above proof) is the following. If the mutation rate is much larger than the migration rates, two genes in different subpopulations are likely to
have undergone a mutation by the time they get together into a single colony, regardless of the precise structure of the population. Hence in that case, $F_{ST}$ depends largely on the probability that two genes in the same subpopulation coalesce before either of them migrates or mutates, which depends mainly on the mutation rate if this is much larger than the migration rates; the dependence of this probability on the different coalescence rates, i.e. on the relative subpopulation sizes, cancels out under weighting by size. At small mutation rates, on the contrary, the probability that two genes in different subpopulations are identical by descent depends strongly on the time it takes them to get together into a single subpopulation and hence on the details of population structure.

We have illustrated in Chapter 4 that the asymptotic behaviour of $F_{ST}$ at very large mutation rates, established above under weighting by size, does not hold under equal weighting of the subpopulations when subpopulation sizes are unequal.

5.3 The mean coalescence time of two individuals from a single subpopulation.

In Section 1 it was proved that under weighting by size, the mean coalescence time of two individuals from the same subpopulation is independent of the migration rates and pattern and, in the time-scaling of this work, equal to $c$. We recall that under the Wright-Fisher model of reproduction within subpopulations, a continuous time-span of length $c$ corresponds to a number of generations equal to the total number of (haploid) individuals in the population. In Chapter 3 it was observed that under various models of population structure the mean coalescence time of two individuals from any single subpopulation is equal to $c$. The population structures for which this was the case all exhibited a lot of "symmetry". In this section, population structure is related to graph theory and we examine under what symmetry and regularity conditions $ET_{ii} = c$ for all $i \in S$.

The lemma below follows immediately from identity (ii) in Theorem 5.1 and from the fact that $\sum_{i \in S} P_i = 1$. In identity (ii) of Theorem 5.1 it is required that the mean coalescence time of every pair of individuals is finite. A sufficient condition (which is in fact necessary - see Theorem 5.5 below) for this to be true was given in Theorem 2.3 in terms of the migration chain of the population, which is the continuous-time Markov chain with Q-matrix $(M_{ij})_{i,j \in S}$, where $M_{ii} = -M_i$ for every $i \in S$.

Lemma 5.4 Assume the population is divided into a finite number of subpopulations and assume the migration chain is irreducible. If the mean coalescence time $ET_{ii}$ of a pair of individuals from subpopulation $i$ does not depend on $i \in S$, then

$$\forall i \in S : ET_{ii} = c.$$
In this section we will give two sets of conditions under which the mean coalescence time of two individuals from the same subpopulation does not depend on the specific subpopulation and is hence equal to \( c \). Both sets of conditions are closely related to symmetry and regularity conditions on graphs. A (simple) graph is an ordered pair consisting of a (countable) set of vertices and a set of edges, each edge being specified by an unordered pair of vertices (see, for example, Bollobás 1979). Equivalently, a graph is determined by a set \( V = \{ v_i \} \) of vertices and an adjacency matrix \( (a_{ij}) \) whose entries are

\[
a_{ij} = \begin{cases} 
1 & \text{if } \{v_i, v_j\} \text{ is an edge} \\
0 & \text{otherwise.}
\end{cases}
\]

In the remainder of this chapter we similarly define a population structure \( \Gamma \) as a triplet \( \Gamma \equiv (S, (c_i)_{i \in S}, (M_{ij})_{i,j \in S}) \) consisting of a (countable) set of subpopulation labels, \( S \), an ordered set of relative subpopulation sizes, \( (c_i)_{i \in S} \), and a matrix of scaled migration rates, \( (M_{ij})_{i,j \in S} \). Considering the subpopulations as vertices and drawing edges between subpopulations if there is gene flow between them, every population structure naturally induces a graph:

Definition 5.1 The induced graph of a population structure \( \Gamma \equiv (S, (c_i)_{i \in S}, (M_{ij})_{i,j \in S}) \) is the graph with vertex-set \( S \) and where \( v_i, v_j \in S : \{i, j\} \) is an edge if and only if \( M_{ij} + M_{ji} > 0 \).

In the theorem below, the assumption of irreducibility of the migration chain, made in Lemma 5.4, is translated in terms of the induced graph. It was seen in Theorem 2.3 that if the number of subpopulations is finite, this assumption is a sufficient condition to ensure that every finite sample of individuals has with probability one a common ancestor within a finite time. We will show here that this assumption is also a necessary condition. For the theorem below to be valid it is of crucial importance that migration is conservative (Assumption 5.1), as has been assumed throughout this chapter.

A graph is said to be connected if for every two vertices \( v_i \) and \( v_j \) there exist \( k \in \mathbb{N} \) and a sequence of vertices

\[
v_i = u_0, u_1, \ldots, u_k = v_j
\]

such that \( u_{l-1} \) and \( u_l \) are adjacent for \( l = 1, \ldots, k \). The sequence (5.6) is called a walk joining \( v_i \) and \( v_j \); \( k \) is the length of the walk (Biggs 1974).

Theorem 5.5 Assume the population is divided into a finite number of subpopulations. Then the following conditions are equivalent:
We show that also $\forall u \in S: P\{T_{ij} < \infty\} = 1$.

Assumption 5.1 yields $c_{ij}, M_{ij}, A_{ij} = c_{ij}, M_{ij}$.

Using (5.10), the left-hand side of this equation can be written as

$$\sum_{u \in V_i} \sum_{v \in S, v \neq u} c_{uv} M_{uv} = \sum_{u \in V_i} \sum_{v \in S, v \neq u} c_{uv} M_{uv}.$$

Proof: It was proved in Theorem (2.3) that (iii) $\Rightarrow$ (ii), while it is evident that (ii) $\Rightarrow$ (i). To prove the theorem, we show that (i) $\Rightarrow$ (iv) and that (iv) $\Rightarrow$ (iii).

(iv) $\Rightarrow$ (iii) As the migration chain has Q-matrix $(M_{ij})_{i,j \in S}$, denoting by $M^{(0)}$ the matrix whose entries are

$$M^{(0)}_{ij} := \begin{cases} M_{ij} & \text{if } j \neq i \\ 0 & \text{if } j = i \end{cases}$$

for $i,j \in S$, irreducibility of the migration chain means that

$$\forall i, j \in S, \exists k \in N : (M^{(0)}_k)_{ij} > 0.$$

Denoting by $M'$ the matrix whose entries are

$$M'_{ij} := \begin{cases} M_{ij} + M_{ji} & \text{if } j \neq i \\ 0 & \text{if } j = i \end{cases}$$

for $i,j \in S$ (cf. Definition 5.1), connectedness of the induced graph means that

$$\forall i, j \in S, \exists k \in N : (M^k)_{ij} > 0.$$

We prove that (5.9) $\Rightarrow$ (5.7).

Assume (5.9) holds and fix $i, j \in S$. We define the set

$$V_i := \left\{ u \in S \mid \exists k \in N : (M^{(0)}_k)_{iu} > 0 \right\}.$$

We must prove that $j \in V_i$. First note that if $u \in V_i$ and $M_{uv} > 0$ then also $v \in V_i$: if $(M^{(0)}_k)_{iu} > 0$ for some $k \in N$ and $M_{uv} > 0$, then $(M^{(0)}_{k+1})_{iv} \geq (M^{(0)}_k)_{iu} M_{uv} > 0$. Hence

$$\forall u \in V_i, \forall v \in S \setminus V_i : M_{uv} = 0. \quad (5.10)$$

We show that also

$$\forall u \in V_i, \forall v \in S \setminus V_i : M_{uv} = 0. \quad (5.11)$$

Assumption 5.1 yields

$$\sum_{u \in V_i} \sum_{v \in S, v \neq u} c_{uv} M_{uv} = \sum_{u \in V_i} \sum_{v \in S, v \neq u} c_{uv} M_{uv}. \quad (5.12)$$

Using (5.10), the left-hand side of this equation can be written as

$$\sum_{u \in V_i} \sum_{v \in S, v \neq u} c_{uv} M_{uv} = \sum_{u \in V_i} \sum_{v \in V_i, v \neq u} c_{uv} M_{uv}$$

$$= \sum_{v \in V_i} \sum_{u \in V_i, u \neq v} c_{uv} M_{uv} \quad (5.13)$$

$$= \sum_{u \in V_i} \sum_{v \in V_i, v \neq u} c_{uv} M_{uv} \quad (5.14)$$
where (5.13) was obtained by swapping the summation indices, and where we have subsequently changed the order of summation to obtain (5.14). Substituting this into equation (5.12) gives

$$\sum_{u \in V_i} \sum_{v \in S \setminus V_i} c_u M_{uv} = 0,$$

so that (5.11) holds. Combining (5.10) and (5.11) we obtain that

$$\forall u \in V_i, \forall v \in S \setminus V_i : M'_{uv} = 0. \quad (5.15)$$

Now be $k \in \mathbb{N}$ such that $(M^{(k)})_{ij} > 0$ (assumption (5.9)). If $k = 0$ then $j = i \in V_i$ and the proof is completed. Assume $k > 0$. Then there exist $i = u_0, u_1, \ldots, u_k = j \in S$ such that $\prod_{l=1}^{k} M'_{u_l} > 0$. We know that $u_0 = i \in V_i$. If $u_l \in V_i$ for some $l \in \{0, \ldots, k-1\}$ then by (5.15) also $u_{l+1} \in V_i$ because $M'_{u_l} > 0$. By induction, it follows that $j \in V_i$.

(i) $\Rightarrow$ (iv) We prove the contrapositive. If the induced graph is not connected, choose $i, j \in S$ in different connected components (Wilson 1972) of the induced graph. Define $T_{ij}^{(r)} := \inf\{t \geq 0 | \alpha(t) \in \{2e^i + e^j; k \in S\}\}$ when $\alpha(0) = e^i + e^j$ to be the first time at which the ancestors of an individual from colony $i$ and an individual from colony $j$ are first simultaneously in the same colony. Note that $T_{ij} \geq T_{ij}^{(r)}$. So

$$0 \leq P\{T_{ij} < \infty\} \leq P\{T_{ij}^{(r)} < \infty\} = 0,$$

contrary to condition (i).

Many symmetry and regularity conditions of graphs are defined in terms of automorphisms: permutations of the vertices that preserve the edges. We introduce an analogous concept for population structures.

**Definition 5.2** An automorphism of a population structure is a permutation of the subpopulation labels that preserves migration rates and subpopulation sizes, i.e. a permutation $\phi : S \to S$ is an automorphism of the population structure $\Gamma \equiv (S, (c_i), (M_{ij}))$ if and only if

$$\forall i, j \in S : M_{\phi(i)\phi(j)} = M_{ij}$$

and

$$\forall i \in S : c_{\phi(i)} = c_i. \quad (5.16)$$

We denote by $Aut(\Gamma)$ the set of all automorphisms of the population structure $\Gamma$. It is easily seen that $Aut(\Gamma)$, with the operation of composition, is a group.

A permutation $\phi : S \to S$ induces an action on every sample from the population: instead of sampling an individual from subpopulation $i \in S$, an individual is sampled from subpopulation $\phi(i)$. Formally, the permutation $\phi$ of $S$ thus induces a permutation of $N^S$: for $\alpha = (\alpha_i) \in N^S$
we denote by \( \phi(\alpha) \) the element of \( N^{S} \) with components \( (\phi(\alpha))_{i} = \alpha_{\phi^{-1}(i)} \) (\( i \in S \)). We prove that if \( \phi \) is an automorphism of the population structure, the coalescence time of a sample from the population has the same distribution as the coalescence time of the "image sample".

**Theorem 5.6** Assume a population with structure \( \Gamma \) and set \( S \) of subpopulation labels. For every \( \phi \in \text{Aut}(\Gamma) \) and for every \( \alpha \in N^{S} \)

\[
T_{\alpha} \overset{\phi}{\Rightarrow} T_{\phi(\alpha)}.
\]

**Proof:** By a coupling argument. Let \( \phi \in \text{Aut}(\Gamma) \). Consider the continuous-time Markov chain \( \zeta(t) := (\alpha(t), \beta(t)) \) where \( \alpha(t) \) is the structured coalescent and where \( \forall t \geq 0 : \beta(t) := \phi(\alpha(t)) \).

Denote by \( B = (B_{\alpha,\beta}) \) the infinitesimal generator of the process \( \beta(\cdot) \). By construction,

\[
B_{\alpha,\beta} = Q_{\phi^{-1}(\alpha), \phi^{-1}(\beta)}
\]

for every \( \alpha, \beta \in N^{S} \), where \( Q \), given by equation (2.2), is the infinitesimal generator of the structured coalescent \( \alpha(\cdot) \). If \( \beta = \alpha \) then \( \phi^{-1}(\beta) = \phi^{-1}(\alpha) \) and

\[
B_{\alpha,\beta} = Q_{\phi^{-1}(\alpha), \phi^{-1}(\alpha)} = \sum_{i \in S} \left\{ (\phi^{-1}(\alpha))_{i} \frac{M_{i}}{2} + \frac{1}{c_{i}} \left( \phi^{-1}(\alpha)_{i} \right)^{2} \right\}.
\]

Because \( \phi \in \text{Aut}(\Gamma) \), substituting \( i := \phi^{-1}(k) \) in the above sum gives

\[
B_{\alpha,\alpha} = \sum_{k \in S} \left\{ (\phi^{-1}(\alpha))_{\phi^{-1}(k)} M_{\phi^{-1}(k)} 2 + \frac{1}{c_{\phi^{-1}(k)}} \left( \phi^{-1}(\alpha)_{\phi^{-1}(k)} \right)^{2} \right\} = \sum_{k \in S} \left\{ \alpha_{k} \frac{M_{k}}{2} + \frac{1}{c_{k}} \left( \alpha_{k} \right)^{2} \right\} = Q_{\alpha,\alpha}.
\]

For \( \beta = \alpha - \varepsilon^{i} + \varepsilon^{j} \) (\( i, j \in S \) with \( j \neq i \)), we have that \( \phi^{-1}(\beta) = \phi^{-1}(\alpha) - \varepsilon^{\phi^{-1}(i)} + \varepsilon^{\phi^{-1}(j)} \) and

\[
B_{\alpha,\beta} = \left( \phi^{-1}(\alpha) \right)_{\phi^{-1}(i)} M_{\phi^{-1}(i)} 2 \frac{M_{\phi^{-1}(j)} \phi^{-1}(j)}{2} \frac{1}{\alpha_{i} \frac{M_{ij}}{2}} \frac{1}{\frac{1}{c_{i}} \left( \frac{\alpha_{i}}{2} \right)^{2}} = Q_{\alpha,\beta}.
\]

Similarly, if \( \beta = \alpha - \varepsilon^{i} \) (\( i \in S \)) then \( \phi^{-1}(\beta) = \phi^{-1}(\alpha) - \varepsilon^{\phi^{-1}(i)} \) and

\[
B_{\alpha,\beta} = \frac{1}{c_{\phi^{-1}(i)}} \left( \phi^{-1}(\alpha)_{\phi^{-1}(i)} \right)^{2} = \frac{1}{c_{i}} \left( \phi^{-1}(\alpha)_{i} \right)^{2} = Q_{\alpha,\beta}.
\]
Finally, if $\beta \notin \{\alpha\} \cup \{\alpha - i'; i \in S\} \cup \{\alpha - i' + j'; i, j \in S, j \neq i\}$, then $B_{\alpha, \beta} = 0 = Q_{\alpha, \beta}$. Hence $B = Q$ so that the process $\beta(\cdot)$ is a version of the structured coalescent.

Starting the structured coalescent $\alpha(\cdot)$ at $t = 0$ from state $\alpha(0) = \alpha \in N^S$, the process $\beta(\cdot)$ starts, by definition, from state $\phi(\alpha)$. The coalescence time $T_\alpha$ of a sample with configuration $\alpha$ is the time until the structured coalescent $\alpha(\cdot)$ enters into the absorbing set of states $\{\gamma \in N^S; \gamma = 1\}$. As $|\alpha(t)| = |\beta(t)|, \forall t \geq 0$, this time is also the time until the process $\beta(\cdot)$ enters into the set $\{\gamma \in N^S; |\gamma| = 1\}$. Because $\beta(\cdot)$ has the distribution of the structured coalescent, the latter absorption time has the same distribution as the coalescence time $T_{\phi(\alpha)}$ of a sample with configuration $\phi(\alpha)$. Thus the distributions of $T_\alpha$ and $T_{\phi(\alpha)}$ are the same.

One way to meet the condition in Lemma 5.4 that $ET_i$ does not depend on $i$ is simply to require that all subpopulations are identical with respect to size, migration pattern and migration rates. Formally, this intuitive description corresponds to the following definition.

**Definition 5.3** A population structure $\Gamma$ is said to be isotropic if its automorphism group $\text{Aut}(\Gamma)$ acts transitively on the set $S$ of the subpopulation labels, that is, if

$$\forall i, j \in S, \exists \phi \in \text{Aut}(\Gamma) : \phi(i) = j.$$  

According to this definition, a population structure is isotropic if all subpopulations have the same size and the migration model is isotropic as defined by Strobeck (1987). If a population structure is isotropic, its induced graph is vertex-transitive (see Biggs 1974 for a definition of vertex-transitivity). Examples of isotropic structures include the symmetric island model, the stepping-stone models and also the four-population model with geographic barrier, discussed in the previous chapters.

**Theorem 5.7 (Strobeck 1987)** Assume the population is divided into a finite number of subpopulations and assume the induced graph is connected. If the structure of the population is isotropic, then

$$ET_{ii} = c$$

for every $i \in S$.

**Proof:** Fix $i$ and $j \in S$. Because the population structure is isotropic, there exists an automorphism $\phi$ of the population structure such that $\phi(i) = j$. By Theorem 5.6 it follows that $T_{ii} = T_{jj}$. As this holds for every $i$ and $j \in S$, Lemma 5.4 gives $ET_{ii} = c, \forall i \in S$. □

Strobeck's result, Theorem 5.7, states that in a population with finite number of subpopulations, connected induced graph and isotropic structure, the mean coalescence time of a pair
of individuals from any single subpopulation is independent of the migration rate and equal to that of two individuals from a panmictic population of the same total size. An alternative set of conditions under which this is the case is in terms of distance.

In a graph there exists a natural distance on the vertices: the distance $d(v_i, v_j)$ between two vertices $v_i$ and $v_j$ is the length of the shortest walk joining $v_i$ and $v_j$ (Biggs 1974, Bollobás 1979). In terms of the adjacency matrix $A$,

$$d(v_i, v_j) = \min\{k \in \mathbb{N}; (A^k)_{ij} > 0\}$$

(where $\min \emptyset \equiv +\infty$, so that the distance between two vertices is infinite if there does not exist a walk joining these vertices). We define the distance between two subpopulations in a population structure to be their natural distance in the induced graph:

**Definition 5.4** Consider a population structure $\Gamma \equiv (S, (c_i)_{i \in S}, (M_{ij})_{i,j \in S})$. The natural distance between subpopulations $i$ and $j$ ($i, j \in S$) is the number

$$d(i,j) := \min\{k \in \mathbb{N}; (M')^k_{ij} > 0\},$$

where $M'$ is the matrix defined by (5.8).

The natural distance between two individuals (or ancestral lineages) in the population is the natural distance between the subpopulations where they are resident.

Tracing the ancestral lineages of a pair of individuals, each time either one lineage migrates, we have, with probability one, one of the following three possibilities: the distance between the lineages either increases by one, decreases by one, or remains the same (this follows from the triangle inequality for the natural distance on the subpopulations). The lineages can coalesce only at times they are resident in a single subpopulation, i.e. when their distance is zero. If every time one of two lineages migrates, the probability that their distance increases or decreases, respectively, depends on the lineages' locations before migration only through their distance before migration, the distribution of their distance at any time will depend on their locations at time zero only through their distance at time zero. If in addition the coalescence rate is the same in every subpopulation, the distribution of the coalescence time of a pair of individuals will depend on their locations at time zero only through their distance at time zero. In particular, the distribution of the coalescence time of two individuals from a single subpopulation will be independent of the specific subpopulation, so that, under the assumptions of Lemma 5.4, their mean coalescence time will be equal to $c$. This is the idea behind the following theorem.
Theorem 5.8 Consider a population with structure \( \Gamma \equiv \left( S, (c_i)_{i \in S}, (M_{ij})_{i, j \in S} \right) \), where \( S = \{1, \ldots, n\} \) \((n < \infty)\). Assume the induced graph is connected and assume

(i) \( \forall i \in S : c_i = c_1 \)

(ii) both

\[
\sum_{k : d(k, j) = d(i, j) - 1} M_{ik} + \sum_{l : d(i, l) = d(i, j) - 1} M_{jl}
\]

and

\[
\sum_{k : d(k, j) = d(i, j) + 1} M_{ik} + \sum_{l : d(i, l) = d(i, j) + 1} M_{jl}
\]

depend on \( i, j \in S \) only through their distance \( d(i, j) \), i.e. if \( i, j, i', j' \in S \) and \( d(i, j) = d(i', j') = d \) then

\[
\sum_{k : d(k, j) = d - 1} M_{ik} + \sum_{l : d(i, l) = d - 1} M_{jl} = \sum_{k : d(k, j') = d - 1} M_{ik} + \sum_{l : d(i', l) = d - 1} M_{jl} =: A_d
\]

and

\[
\sum_{k : d(k, j) = d + 1} M_{ik} + \sum_{l : d(i, l) = d + 1} M_{jl} = \sum_{k : d(k, j') = d + 1} M_{ik} + \sum_{l : d(i', l) = d + 1} M_{jl} =: B_d.
\]

Then

\( \forall i \in S : ET_{ii} = c \).

Proof: First note that for \( i, k \in S \) with \( k \neq i \), \( M_{ik} > 0 \) implies that \( d(i, k) = 1 \), so that assumption (5.18) for \( d = 0 \) implies that for every \( i, j \in S : M_i = M_j =: M \) \((M = B_0/2)\), i.e. the total scaled migration rate out of each subpopulation is the same.

Consider two individuals in subpopulation \( i \) at time zero and denote by \( D_{ii}(t) \) the natural distance between their ancestral lineages time \( t \) ago, where we define \( D_{ii}(t) := -1 \) if the two individuals share the same ancestor time \( t \) ago. In terms of the structured coalescent \( \{\alpha(t) : t \geq 0\} \), starting from \( \alpha(0) = 2e^t \), we have that

\[
D_{ii}(t) := \begin{cases} 
  d(k, l) & \text{if } \alpha(t) = \epsilon^k + \epsilon^l \\
  -1 & \text{if } |\alpha(t)| = 1.
\end{cases}
\]

From the assumptions (and from the Markov character of the structured coalescent) it follows that the process \( \{D_{ii}(t) : t \geq 0\} \) is Markov with state space \( \mathbb{N} \cup \{-1\} \) and with Q-matrix \( \Delta = (\Delta_{r,s}) \) given by

\[
\Delta_{0,s} = \begin{cases} 
  - \left( \frac{1}{c_1} + M \right) & \text{if } s = 0 \\
  \frac{1}{c_1} & \text{if } s = -1 \\
  M & \text{if } s = 1 \\
  0 & \text{otherwise},
\end{cases}
\]
Chapter 5: GENERAL RESULTS

\[
\Delta_{r,s} = \begin{cases} 
-\frac{1}{2}(A_r + B_r) & \text{if } s = r \\
 \frac{1}{2}A_r & \text{if } s = r - 1 \\
 \frac{1}{2}B_r & \text{if } s = r + 1 \\
 0 & \text{otherwise}
\end{cases}
\]

\( (r \in \mathbb{N}_0) \), and \( \Delta_{r-1,s} = 0 \) for all \( s \). As this Q-matrix does not depend on \( i \in \mathcal{S} \) and as \( D_{ii}(0) = 0 \), independent of \( i \in \mathcal{S} \), the distribution of \( D_{ii}(t) \) is for every \( t \geq 0 \) independent of \( i \). Because

\[
P\{T_{ii} \leq t\} = P\{D_{ii}(t) = -1\}
\]

it follows that the distribution of the coalescence time \( T_{ii} \) of two individuals from subpopulation \( i \) does not depend on \( i \). Hence Lemma 5.4 yields \( ET_{ii} = c \), \( \forall i \in \mathcal{S} \). \( \Box \)

The assumptions of Theorem 5.8 are closely related to the graph-theoretic concept of distance-regularity. A distance-regular graph is a connected graph with the following property: for every \( d \in \mathbb{N} \) there exist numbers \( a_d \) and \( b_d \in \mathbb{N} \) such that for each pair \((u, v)\) of vertices satisfying \( d(u, v) = d \) it holds that

(i) the number of vertices adjacent to \( u \) and at distance \( d - 1 \) from \( v \) is \( a_d \)

(ii) the number of vertices adjacent to \( u \) and at distance \( d + 1 \) from \( v \) is \( b_d \)

(Biggs 1974). Note that a distance-regular graph is always regular, that is, for every vertex, the number of vertices adjacent to it is the same, namely \( b_0 \) (Biggs 1974). The condition of distance-regularity provides the following special case of Theorem 5.8.

**Corollary 5.9** Consider a population with structure \( \Gamma \equiv (S, (c_i)_{i \in \mathcal{S}}, (M_{ij})_{i,j \in \mathcal{S}}) \), where \( S = \{1, \ldots, n\} \) \((n < \infty)\). Assume that

(i) \( \forall i \in \mathcal{S} : c_i = c_1 \)

(ii) the induced graph is distance-regular

(iii) for every \( i, j \in \mathcal{S} \) with \( j \neq i \): \( M_{ij} = M_{ji} \in \{0, \frac{M}{b_0}\} \) \((\text{where } b_0 \text{ is as in the above definition of distance-regularity), i.e. migration between subpopulations is symmetric and all non-zero migration rates are equal.} \)

Then

\[ \forall i \in \mathcal{S} : ET_{ii} = c. \]

**Proof:** With \( a_d \) and \( b_d \) as in the definition of distance-regularity (assumed to hold for the induced graph), we have for every \( i, j \in \mathcal{S} \) with \( d(i, j) = d \) that

\[
\sum_{k \colon d(k, j) = d-1} M_{ik} + \sum_{l \colon d(i, l) = d-1} M_{jl} = 2a_d \frac{M}{b_0} =: A_d
\]
and

$$\sum_{k:d(k,j)=d+1} M_{ik} + \sum_{i:d(i,i)=d+1} M_{ji} = 2b_d \frac{M}{b_0} =: B_d,$$

independent of $i$ and $j$ with $d(i,j) = d$. Hence the conditions of Theorem 5.8 are satisfied. □

Theorems 5.7 and 5.8 state two sets of conditions under which the mean coalescence time of two individuals from any single subpopulation is independent of the migration rate and equal to that of two individuals from a panmictic population of the same total size. One may ask about the relationship between these two sets of conditions.

We first note that both sets of conditions require that

(i) all subpopulations have the same size
(ii) the induced graph is regular
(iii) the total scaled migration rate from each subpopulation is the same (i.e. all $M_i$ are equal).

Statements (i) and (iii) are obvious. From the condition of isotropy (Theorem 5.7), statement (ii) is immediate, as every vertex-transitive graph is regular (Biggs 1974). Similarly, (ii) follows immediately from the conditions of Corollary 5.9, as every distance-regular graph is regular. To see that the conditions of Theorem 5.8 also require (ii), first note that condition (5.17) for $d = 1$ yields that $M_{ij} + M_{ji} = A_1$ for every $i,j \in S$ with $d(i,j) = 1$. As all subpopulations have the same size, Assumption 5.1 gives that $M_{1i} = \sum_{j \neq i} M_{ij}$ for every $i \in S$. Hence, denoting by $k_i$ the valency of subpopulation $i$, that is, the number of subpopulations adjacent to subpopulation $i$ in the induced graph, we have that

$$M_i = \sum_{j \neq i} M_{ij} = \frac{1}{2} \sum_{j \neq i}(M_{ij} + M_{ji}) = \frac{1}{2} k_i A_1.$$

By condition (5.18) for $d = 0$, all $M_i$ are equal ($M_i = B_0/2$, $\forall i \in S$). Hence all $k_i$ are equal ($k_i = B_0/A_1$, $\forall i \in S$), i.e. all subpopulations have the same valency in the induced graph, which means that the induced graph is regular.

These observations are helpful in recognizing population structures that do not meet the conditions of Theorems 5.7 or 5.8. For example, the continental island model (where the induced graph is not regular) and the two-population model with unequal subpopulation sizes do not satisfy either set of conditions. Indeed, we calculated in Chapter 3 that in the two-population model, the mean coalescence time of two individuals from the larger subpopulation is always larger than $c$, while that of two individuals from the smaller subpopulation is smaller than $c$. In a continental island model with at least two islands ($n \geq 3$), the mean coalescence time of two individuals from the continent is always larger than $c$ and that of two individuals from a single island is always smaller than $c$, regardless of the relative sizes of continent and islands. Another example of a
population structure whose induced graph is not regular is a stepping-stone model on a finite line (see figure 5.1). For \( n = 5 \) subpopulations, we have checked that under this model, the mean coalescence time of two individuals from a single subpopulation is larger as the subpopulation is more central and smaller as the subpopulation is situated closer to the ends of the line (see also Hey 1991). Recall that if one takes the average over all the subpopulations, weighting each subpopulation by its relative size, the mean coalescence time of two individuals from the same subpopulation is equal to \( c \), for all these population structures (Theorem 5.1.ii).

Among the models of population structure described in the previous chapters, the symmetric island model and the circular stepping-stone model meet both sets of conditions: those of Theorem 5.7 and those of Corollary 5.9. It is easily seen that a stepping-stone model on a \( K \times L \) torus, with \( K, L \geq 2 \) and \( \max(K, L) \geq 3 \), and the four-population model with geographic barrier (\( \epsilon \neq \frac{1}{2} \)) do not meet the conditions of Theorem 5.8. Yet for both these models the conditions of Theorem 5.7 are satisfied and \( \mathbb{E} T_{ii} = c, \forall i \in S \). It is more difficult to find an example of a population structure meeting the conditions of Theorem 5.8 and not those of Theorem 5.7. Adel'son-Velskii (1969) have found an example of a distance-regular graph which is not vertex-transitive (see also Biggs 1974, p.139, 20D): the graph \( \Psi \) whose vertices are the 26 symbols \( x_i, y_i \) (\( i \in \mathbb{Z}_{13} \)) and where

\[
\begin{align*}
x_i \text{ and } x_j \text{ are adjacent } & \iff |i - j| = 1, 3, 4 \\
y_i \text{ and } y_j \text{ are adjacent } & \iff |i - j| = 2, 5, 6 \\
x_i \text{ and } y_j \text{ are adjacent } & \iff i - j = 0, 1, 3, 9
\end{align*}
\]

is distance-regular (with \( b_0 = 10, b_1 = 6, a_1 = 1 \) and \( a_2 = 4 \)) but is not vertex-transitive. Placing equal-sized subpopulations at the vertices of this graph (\( S = \{ x_i ; i \in \mathbb{Z}_{13} \} \cup \{ y_i ; i \in \mathbb{Z}_{13} \} \)) and assigning a migration rate of \( \frac{M}{10} \) between every pair of adjacent vertices, in both directions (i.e. for \( u, v \in S \) with \( u \neq v \) : \( M_{uv} = \frac{M}{10} \) if \( u \) and \( v \) are adjacent in \( \Psi \), \( M_{uv} = 0 \) otherwise), a population structure is obtained which satisfies the conditions of Corollary 5.9 but which is not isotropic, as its induced graph is not vertex-transitive. Hence the conditions of Theorems 5.7 and 5.8 are different, and neither implies the other.

---

Figure 5.1: The finite linear stepping-stone model. This model assumes a finite number, \( n \), of subpopulations (in this picture: \( n = 5 \)), equal-sized subpopulations and symmetric nearest neighbour migration: \( S = \{ 1, \ldots, n \} \), \( c_i = 1 \) for \( i = 1, \ldots, n \) and \( M_{ij} = M/2 \) if \( |i - j| = 1 \), \( M_{ij} = 0 \) otherwise (\( i, j = 1, \ldots, n; j \neq i \)).
Chapter 5: GENERAL RESULTS

Figure 5.2: Example of a model of population structure under which the mean coalescence time of two individuals from any single subpopulation is equal to that of two individuals from a panmictic population of the same total size, but which meets the conditions of neither Theorem 5.7 nor Theorem 5.8.

We also note that the conditions given in Theorems 5.7 and 5.8 are sufficient but not necessary for the mean coalescence time of two individuals from any single subpopulation to be equal to \( c \). For example, the population structure with \( n = 3 \) subpopulations, \( c_1 = 1 \) and \( c_2 = c_3 = 2 \), and \( M_{12} = M_{13} = 6M \), \( M_{21} = M_{31} = 3M \) and \( M_{23} = M_{32} = M \) (see figure 5.2), meets neither set of conditions, yet satisfies

\[
ET_{ii} = 5 \text{ for } i = 1, 2, 3.
\]

However, in this model of population structure, the \textit{distribution} of the coalescence time of two individuals from a single subpopulation depends on the subpopulation:

\[
E[e^{-sT_{11}}] = \frac{450M^3 + 300M^2s + 42Ms^2 + 43 + 42M^2 + 23Ms + 2s^2}{\Omega}
\]

\[
E[e^{-sT_{22}}] = \frac{450M^3 + 300M^2s + 47Ms^2 + 23 + 42M^2 + 23Ms + 2s^2}{\Omega} = E[e^{-sT_{33}}]
\]

where

\[
\Omega = 2250M^3s + 900M^2s^2 + 110M s^3 + 4s^4 + 450M^3 + 510M^2s + 109M s^2 + 6s^3 + 42M^2 + 23Ms + 2s^2.
\]

The results described in this section are specific to samples of size two. It was seen in Chapter 3 that under the symmetric island model (which meets the conditions of both Theorems 5.7 and 5.8), the mean time since the most recent common ancestor of three individuals from a single subpopulation \( (ET_{3e_1}, \text{ given by equation (3.14)}) \) depends on the migration rate and is larger than that of three individuals from a panmictic population of the same total size. The mean time since three individuals from a single subpopulation descended from two distinct ancestors \( (ET_{3e_2}, \text{ given by equation (3.12)}) \) also depends on the migration rate and is smaller than in the panmictic case. Both \( ET_{3e_1} \) and \( ET_{3e_2} \) converge, as the migration rate \( M \) tends to infinity, to the corresponding mean times in a panmictic population of the same total size.
5.4 The $F_{ST}$ value of a pair of neighbouring subpopulations.

The quantity $F_{ST}$ is commonly used by population geneticists to estimate the effective level of gene flow among subpopulations, defined as

\[(N_m)_e = \frac{1}{4} \left( \frac{1}{F_{ST}} - 1 \right). \tag{5.19} \]

For a diploid population under an island model of population structure (and the neutral Wright-Fisher model of reproduction), with a large number of subpopulations and a small mutation rate, \((N_m)_e\) is approximately the number of migrants (i.e. half the number of migrant genes) per subpopulation per generation. However, it was shown in Chapter 4 that the relationship between $F_{ST}$ and the actual migration rate varies considerably among population structures, so that \((N_m)_e\) provides at most a vague idea of the actual level of gene flow between subpopulations. In this section we prove that the relationship between the actual migration rate and the $F_{ST}$ value of a pair of neighbouring subpopulations is the same for a range of (sufficiently symmetric) models of population structure, in the limit as the mutation rate tends to zero. This uniform relationship might give rise to better methods for inferring information on the actual level of gene flow between subpopulations.

**Theorem 5.10** Assume

(i) \( \forall i \in S : c_i = 1 \) and \( M_i = M \)

(ii) the distribution of the coalescence time of two individuals from the same subpopulation does not depend on the subpopulation; say \( T_i \overset{\Delta}{=} T_0, \forall i \in S \)

(iii) the distribution of the coalescence time of two individuals from neighbouring subpopulations does not depend on the specific subpopulations; say \( T_{ij} \overset{\Delta}{=} T_1, \forall i, j \in S \) with \( d(i, j) = 1 \).

Then, denoting by $F_{ST}(d = 1)$ and $F_{ST}^{(0)}(d = 1)$ the values of $F_{ST}$ and $F_{ST}^{(0)}$, respectively, of a pair of subpopulations at distance one,

\[
F_{ST}(d = 1) = \frac{1}{1 + 2M} \frac{1 - F_0}{1 - (1 + 2 \theta) F_0} \tag{5.20}
\]

\[
F_{ST}^{(0)}(d = 1) = \begin{cases} 
\frac{1}{1 + 2M} \frac{ET_0}{ET_0 - 1} & \text{if } ET_0 < \infty \\
\frac{1}{1 + 2M} & \text{if } ET_0 = \infty
\end{cases} \tag{5.21}
\]

where, as in Chapter 4, \( f_0 \) is the probability of identity by descent of two genes from the same subpopulation and \( \theta \) is the scaled mutation rate.

**Proof:** As all subpopulations have the same size, two genes randomly sampled from the union of two subpopulations are with probability 1/2 from the same subpopulation and with probability
1/2 from different subpopulations. Hence, calculating the $F_{ST}$ value of a pair of neighbouring subpopulations, the probability of identity by descent of two genes sampled from the union of these subpopulations is

$$f(d = 1) = \frac{f_0 + f_1}{2}$$

(where $f_1$ denotes the probability of identity by descent of two genes at distance one), so that

$$F_{ST}(d = 1) = \frac{f_0 - f_1}{2f_0 - f_1} = \frac{1}{1 + 2\frac{1-f_0}{f_0 - f_1}}. \quad (5.22)$$

In the system of equations, (2.44), for the moment-generating function of the coalescence time of a sample from the population, the equation corresponding to $\alpha = 2e^t$ is for every $i \in S$ and for $s = \theta$ given by

$$(1 + M + \theta)f_0 - Mf_1 = 1,$$

so that

$$f_0 - f_1 = \frac{1 - (1 + \theta)f_0}{M}. \quad (5.23)$$

Substituting this result into (5.22), equation (5.20) is obtained.

To prove equation (5.21), we calculate the limit of $F_{ST}(d = 1)$ as $\theta \downarrow 0$. If $ET_0 < \infty$ then, by the monotone convergence theorem,

$$\lim_{\delta \downarrow 0} f_0 = P{T_0 < \infty} = 1$$

and, applying l'Hopital's rule,

$$\lim_{\delta \downarrow 0} \frac{1 - f_0}{1 - (1 + \theta)f_0} = \frac{ET_0}{ET_0 - 1}.$$  

Hence in that case,

$$F_{ST}^{(0)}(d = 1) = \lim_{\delta \downarrow 0} F_{ST}(d = 1) = \frac{1}{1 + 2M \frac{ET_0}{ET_0 - 1}}.$$  

If $ET_0 = \infty$ and $P{T_0 < \infty} = 1$, l'Hopital's rule gives

$$\lim_{\delta \downarrow 0} \frac{1 - f_0}{1 - (1 + \theta)f_0} = \lim_{\delta \downarrow 0} \frac{E[T_0 e^{-\theta T_0}]}{(1 + \theta)E[T_0 e^{-\theta T_0}] - f_0} = 1,$$

so that

$$F_{ST}^{(0)}(d = 1) = \lim_{\delta \downarrow 0} \frac{1}{1 + 2M \frac{1-f_0}{f_0 - f_1}} = \frac{1}{1 + 2M}.$$  

If $\lim_{\delta \downarrow 0} f_0 = P{T_0 < \infty} < 1$, then

$$\lim_{\delta \downarrow 0} \frac{1 - f_0}{1 - (1 + \theta)f_0} = \frac{1 - P{T_0 < \infty}}{1 - P{T_0 < \infty}} = 1,$$
so that also in this case
\[ F^{(0)}_{ST}(d = 1) = \frac{1}{1 + 2M}. \]

Result (5.21) with \( E\bar{T}_0 = \infty \) applies to the infinite stepping-stone models in one and two dimensions, discussed in Chapters 3 and 4. The infinite stepping-stone model in three dimensions (a three-dimensional rectangular lattice of equal-sized subpopulations with an infinite number of subpopulations in all three dimensions, where every gene has a scaled migration rate of \( M/12 \) to each of the six neighbouring subpopulations) also meets the conditions of Theorem 5.10, with \( E\bar{T}_0 = \infty \) because the symmetric random walk in three dimensions is transient. Hence also under the infinite three-dimensional stepping-stone model,
\[ F^{(0)}_{ST}(d = 1) = \frac{1}{1 + 2M}. \]

For the case of a finite number of subpopulations, some conditions under which assumptions (i) and (ii) of Theorem 5.10 are fulfilled were set out in the previous section. Combined with these conditions, result (5.21) with \( E\bar{T}_0 < \infty \) leads to Corollary 5.11. In that Corollary we also use the following definition.

**Definition 5.5** A population structure \( \Gamma \equiv (S, (c_i)_{i \in S}, (M_{ij})_{i,j \in S}) \) is said to be edge-transitive if
\[
\forall i, j, k, l \in S \text{ with } d(i, j) = d(k, l) = 1, \exists \phi \in Aut(\Gamma) : \{\phi(i), \phi(j)\} = \{k, l\}.
\]

Note that edge-transitivity of a population structure does not imply isotropy. For example, the two-population model with unequal subpopulation sizes \((c_1 \neq c_2)\) and the continental island model are edge-transitive but not isotropic. The stepping-stone model on the torus with \( K, L \geq 2 \) and the four-population model with barrier \((\epsilon \neq 1/2)\) are isotropic but not edge-transitive.

**Corollary 5.11** Assume the population is divided into \( n \) subpopulations \((S = \{1, \ldots, n\}, n < \infty)\) and assume its induced graph is connected. If \( c_1 = 1, M_1 = M \) and either

(a) the population structure is isotropic and edge-transitive

or

(b) the population structure satisfies the conditions of Theorem 5.8, then
\[ F^{(0)}_{ST}(d = 1) = \frac{1}{1 + 2M \frac{n}{n-1}}. \]
Proof: It was noted earlier that both the condition of isotropy and the conditions of Theorem 5.8 imply equal subpopulation sizes and the same migration rate from each subpopulation. Hence \( c_i = 1 \) and \( M_i = M \), for every \( i \in S \). Under assumption (a), it follows from Theorem 5.6 and the definition of isotropy that

\[
\forall i, j \in S: T_{ij} \overset{d}{=} T_{jj},
\]

while Theorem 5.6 and the definition of edge-transitivity yield

\[
\forall i, j, k, l \in S \text{ with } d(i, j) = d(k, l) = 1 : T_{ij} \overset{d}{=} T_{kl}.
\]

If assumption (b) holds, it was shown in the proof of Theorem 5.8 that the distribution of \( T_{ii} \) does not depend on \( i \in S \). The same argument can in fact be used to prove that under the conditions of Theorem 5.8, the distribution of the coalescence time of any pair of individuals depends on their locations only through their distance. In particular, the distribution of the coalescence time of a pair of individuals from neighbouring subpopulations does not depend on the pair of neighbouring subpopulations they are taken from. Hence, both under assumptions (a) and (b), the conditions of Theorem 5.10 are satisfied and equation (5.21) holds, where \( ET_0 = n \) by Theorems 5.7 or 5.8, respectively. \( \square \)

Examples of population structures that are both isotropic and edge-transitive include the symmetric island model, the circular stepping-stone model and a stepping-stone model on a square torus \( (K = L) \).

If it is reasonable to make the assumptions of Theorem 5.10 or Corollary 5.11, the results proved in this section may be used to estimate the actual migration rate between subpopulations, rather than the effective level of gene flow estimated according to equation (5.19). For example, under the assumptions of Corollary 5.11,

\[
M = \frac{1}{2} \left(1 - \frac{1}{n}\right) \left(\frac{1}{F_{ST}^{(0)}(d = 1)} - 1\right) = \frac{n - 1}{ET_1 - n}.
\]

Alternatively, it follows from equation (5.23) that

\[
M = \frac{1 - (1 + \theta)f_0}{f_0 - f_1}
\]

under the assumptions of Theorem 5.10. In addition, an estimate of the fraction \( \frac{M}{4(Nm)_{eq}} \) of the actual over the effective migration rate may provide information on the underlying population structure. For example, if the symmetric island model is an appropriate description of the structure of the population, \( \frac{M}{4(Nm)_{eq}} \) should be smaller than one (under this model, \( \frac{M}{4(Nm)_{eq}} \) equals one in the limit as \( n \to \infty \) and \( \theta \downarrow 0 \)), while a value of \( \frac{M}{4(Nm)_{eq}} \) significantly larger than one suggests isolation by distance.
REFERENCES

ADEL’SON-VELSKII, G. M., 1969 Example of a graph without a transitive automorphism group, 

BIGGS, N., 1974 Algebraic Graph Theory. Cambridge University Press.


CANNINGS, C., 1974 The latent roots of certain Markov chains arising in genetics: A new 

CHAKRABORTY, R., and DANKER-HOPFE, H., 1991 Analysis of population structure: A 
comparative analysis of different estimators of Wright’s fixation indices. In Handbook of 
Chapter 7: pp. 203-254.


Wiley, New York.


FELLER, W., 1966 An Introduction to Probability Theory and Its Applications, Vol. II. Wiley, 
New York.


Press, Oxford.

GRIFFITHS, R. C., 1980 Lines of descent in the diffusion approximation of neutral Wright-Fisher 

GRIFFITHS, R. C., 1981 The number of heterozygous loci between two randomly chosen com-
pletely linked sequences of loci in two subdivided population models. J. Math. Biol. 12: 
251-261.


