UNIVERSITY OF CALIFORNIA, MERCED

MASTER'S THESIS

Genetic Correlation in the One-Dimensional Stepping Stone Model of Population Structure

by

Elizabeth Owens

A technical report submitted
in partial fulfillment of the requirements for the degree of

Master of Science in Applied Mathematics

2015

Committee Members:
Professor Suzanne Sindi, Chair
Professor Arnold Kim
Professor Karin Leiderman

UNIVERSITY OF CALIFORNIA, MERCED
Graduate Division

This is to certify that I have examined a copy of a technical report by

Elizabeth Owens

and found it satisfactory in all respects, and that any and all revisions required by the examining committee have been made.

Applied Mathematics
Graduate Studies Chair:

_____
Professor Boaz Ilan

Thesis Committee:

_____
Professor Arnold Kim

Thesis Committee:

_____
Professor Karin Leiderman

Committee Chair / Research Advisor:

_____
Professor Suzanne Sindi

_____
Date

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Suzanne Sindi. Her support and enthusiasm have been a constant throughout this project and my entire graduate school experience.

Furthermore, I would like to thank the members of my committee, Professors Arnold Kim and Karin Leiderman, for their input and feedback. Thank you also to my friend and colleague Jason Davis, whose technological expertise was invaluable.

Last, but certainly not least, I would like to thank my family. Without your love, I would not be where I am today.

Contents

Genetic Correlation in the One-Dimensional Stepping Stone Model of Population Structure

by

Elizabeth Owens

Master of Science in Applied Mathematics

Dr. Suzanne Sindi, Committee Chair

University of California, Merced

2015

Abstract

We explore the relationship between physical distance and genetic correlation. We focus on the one-dimensional stepping-stone model of population structure, which describes the evolution of a neutral allele in a population that has been subdivided into a number of discrete islands. The generational processes of migration and reproduction are simulated for this population, and we investigate how these forces impact r_k, the correlation between islands at a distance k. We consider different geographic structures - linear and circular arrangements of islands - as well as different migration patterns. We compare our results with asymptotic results derived by Kimura and Weiss under the assumption of infinitely many islands. We find substantial deviation from these asymptotic results especially with regard to long-distance migration.

List of Symbols

- n: Number of islands
- N: Number of individuals per island
- p_i: Frequency of desired allele on island i
- m_j: Rate of migration to islands j steps away
- m_∞: Rate of "long-range" migration (mixture of the full population)
- r_k: Genetic correlation coefficient for islands that are k steps apart

List of Figures

Chapter 1

Introduction

1.1 Population Genetics

Population genetics concerns itself with the evolution of the genetic composition of groups of individuals. Over the past century and a half, the field has experienced a rapid evolution of its own. Since Darwin's "On the Origin of Species" in 1859 and the introduction of Mendelian inheritance in the 1860s, contributions from many different groups and individuals have advanced our understanding to its current state.

Typically, when discussing the genetic composition of a population we are concerned with the frequency of certain genetic information in that population. Sites within the genome, known as loci, have distinct possibilities for the information encoded at that site. Each possibility is represented by an allele, and a particular organism's combination of alleles make up its genotype (the inheritable information that determines the expression of a certain trait). For simplicity, many mathematical models of population genetics consider a limited number of loci with finitely many alleles. One of the most common models considers a single genomic locus with two alleles. For example, one might consider the locus encoding hair length in cats: the two alleles for this site determine if the cat has long hair or short hair.

Hardy [5] and Weinberg [16] discovered independently that in well-mixed populations with random mating, the frequencies of different genotypes will eventually arrive at equilibrium. For a single locus with two alleles A_1 (occurring with frequency p) and A_2 (with frequency $1 - p = q$), the ratio of the genotypes A_1A_1, A_1A_2, and A_2A_2 becomes $p^2 : 2pq : q^2$ in this equilibrium state. However, it is uncommon for populations to meet the conditions required for exact Hardy-Weinberg equilibrium: perfectly random mating with no other active evolutionary processes. Populations may deviate from equilibrium if they experience inbreeding, mutation, genetic drift, natural selection, or any other influences [4, p. 81].

Two main schools of thought arose to explain the primary underlying cause of deviation from Hardy-Weinberg equilibrium. The first of these stemmed from the work of Darwin, and was based on the idea that natural selection would be the main driving force in allele frequency evolution. Proponents of this theory argued that certain alleles can

provide an evolutionary advantage, so a population should evolve based on "survival of the fittest". While this is reasonable, and definitely one of the factors in population development, selection does not tell the entire story. Natural selection is more prevalent in larger, well-mixed populations. But what can be said for populations that do not meet these criteria?

In contrast to selection, the neutral theory was proposed by Kimura in the 1960s [8], and its introduction caused waves of controversy within the population genetics community. In this theory, the dominant force causing deviation from Hardy-Weinberg equilibrium is due to unbiased genetic drift where an allele can be over-represented in a population simply by random sampling. In the neutral theory, also known as genetic drift, it is possible for every individual in a population to carry a given allele (we say the allele has become fixed) without it providing any advantage in fitness. Alleles are seen as neutral, hence the name of the theory, and their frequency drifts over time due to random changes. Genetic drift has been found to dominate in smaller populations, or populations that are subdivided.

1.2 Population Subdivision

Subdivision of a population into multiple groups or colonies is a significant factor in genetic evolution because of the dramatic effects it can have on subpopulations. If two subpopulations are far enough removed from one another (with little or no mating between them) for a long enough period of time, they can evolve to become two separate species; this process is known as allopatric speciation [2]. It has recently been shown that the rate of formation of new species is closely linked to migration patterns within a subdivided population [19]. Of course, not every case is so drastic as to result in speciation— the concept of "isolation by distance" and its effects are observed on more subtle levels as well.

Wright [18] developed a simple model of a subdivided population to establish the effects of isolation by distance. In this model, the population is divided into an effectively infinite number of equally-sized islands. Individuals may migrate from any given island to another, and each island's current residents breed among themselves. This basic model opened up a new realm of ideas for modeling population structure and gene flow (see Figure 1.1).

The "stepping-stone" models were first developed by Kimura in 1953 [7], and the details were fleshed out more than a decade later by Kimura and Weiss [9] (see Section 2.1). These models extend to multiple dimensions: the one-dimensional model can be used to represent a population that has been subdivided linearly (such as along a coastline or mountain ridge) while the two-dimensional model would represent a population subdivided across a plane (as in a desert or plain). There is also a three-dimensional model that is sometimes used for ocean-dwelling populations that have been subdivided not only on a plane, but also at different depths. Because of their intuitive nature and the fact that they can be applied to so many different populations, the stepping stone models have remained extremely popular. They are often used as a prime example

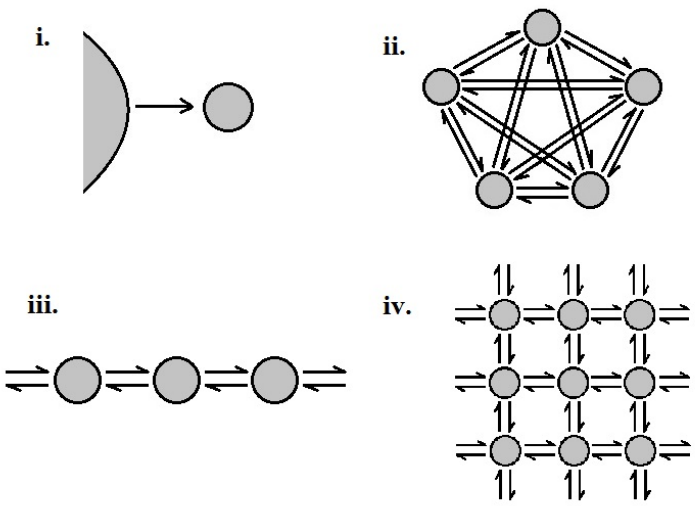


Figure 1.1: Island Models: (i) Continent-island model; (ii) Wright's island model; (iii) One-dimensional stepping stone model; (iv) Two-dimensional stepping stone model. Each island has population size N and total migration rate m.

of isolation by distance, because a greater number of steps between two subpopulations corresponds to a higher level of genetic difference between individuals in those subpopulations.

While Kimura and Weiss laid a solid framework for the stepping-stone models, there have been others who have sought to employ similar methods in more realistic population migration settings. Maruyama [11] considered the more physically realistic condition of finitely many islands. Over the course of several papers, Maruyama explored the one-dimensional stepping stone model with a variety of new conditions, including linear cases where the boundary islands can either "reflect" or "absorb" the immigrants they receive [11], a circular population structure [11], and a case where migration is not symmetric (migration is more likely to occur in one direction than the other) [10].

Through his analysis, Maruyama obtained several key results. First and foremost, by extending his matrix system to infinite dimensions, he was able to verify Kimura and Weiss's formula for the correlation coefficient (see Eq. 2.2). He was able to identify the necessity of having a positive (nonzero) long-range migration term: it "serves mathematically as a stabilizer, and without this term there is no meaningful stationary correlation of the gene frequencies among colonies" [11]. Maruyama also performed the first computer simulation experiments of the stepping-stone model, using a Monte Carlo method [11]. While these simulations verified his analytical results, they were limited in scale: the full population simulated only consisted of 5 or 10 subpopulations.

In the current work, we chose to focus on attributes that better reflect real populations. We developed more extensive computer models to verify and extend the work of Kimura, Weiss, and Maruyama, and we sought to bridge the gaps between simulation and theory. We used systems on the order of 10^3 subpopulations, which are

larger than those of Maruyama while still remaining finite. Through simulation of the original generational processes outlined by Kimura and Weiss, we investigated a linear population structure and compared it to a theoretical asymptotic result. Next, we explored a circular population structure, and compared the results to the linear model. Finally, we imposed a previously unstudied pattern of distributed long-range migration— a pattern with greater biological significance.

Chapter 2

Linear Model

2.1 Introduction

The classic results for the linear model of population structure were established by Kimura and Weiss [9]. We will be considering only the one-dimensional stepping stone model, which consists of an infinite line of discrete colonies (Figure 2.1) evolving over time in discrete non-overlapping generations. Each generation consists of a migration step and a reproduction step. During migration, individuals may migrate one colony to the right or left, each with probability $m_1/2$, so that the total rate of one-step migration is m_1. Alternatively, they may engage in "long-range" migration with likelihood m_∞, where $m_\infty \ll m_1$. This represents the colony exchanging individuals with a random sample taken from the entire population. Once all migration has been completed, individuals may reproduce within their respective islands. Reproduction is modeled by randomly selecting a certain number of individuals to pass on their genetic information (allele value) to the next generation.

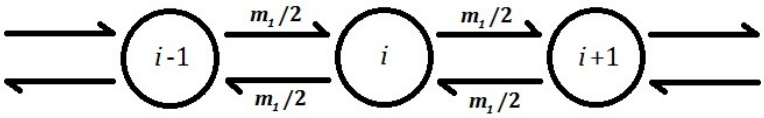


Figure 2.1: The one-dimensional stepping stone model, as described in [9].

There are many ways to quantify the degree of genetic distance in a population, but we chose to focus on the measure used in the original work of Kimura and Weiss: the correlation coefficient r_k [9]. This quantity considers allele frequencies on all pairs of islands that are k steps apart, and calculates a ratio of their covariance and variance:

$$r_k = \frac{E_\phi(\tilde{p}_i\tilde{p}_{i+k})}{V_p} = \frac{E_\phi(\tilde{p}_i\tilde{p}_{i+k})}{E_\phi(\tilde{p}_i^2)} = \frac{E_\phi[(p_i-\bar{p})(p_{i+k}-\bar{p})]}{E_\phi[(p_i-\bar{p})^2]}. \tag{2.1}$$

Note that p_i is the allele frequency on island i at the designated time, $\bar{p}$ is the average allele

frequency among all islands, and E_ϕ is the expectation of gene frequencies among colonies.

This statistic r_k bears similarity to r, the Pearson product-moment correlation coefficient, which is defined for two variables X and Y and sample size n by

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Here the variables in question are the allele frequencies in the ith and $i+k$th islands, so the mean value for each one is the same as that of the entire population ($X_i = p_i$ and $Y_i = p_{i+k}$, so $\bar{X} = \bar{Y} = \bar{p}$). Substituting this in, we see that we obtain the same form as Kimura and Weiss's r_k:

$$r_k = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(p_{i+k} - \bar{p})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2}\sqrt{\sum_{i=1}^{n}(p_{i+k} - \bar{p})^2}} = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(p_{i+k} - \bar{p})}{\sum_{i=1}^{n}(p_i - \bar{p})^2}.$$

Kimura and Weiss found that when the population has reached a steady state, the correlation coefficient r_k exhibits exponential decay as the physical distance k between islands increases (Figure 2.2). The rate of decay is determined by the migration rates, and r_k is calculated in the one-dimensional case by

$$r_k = e^{-\sqrt{\frac{2m_\infty}{m_1}}k}. \tag{2.2}$$

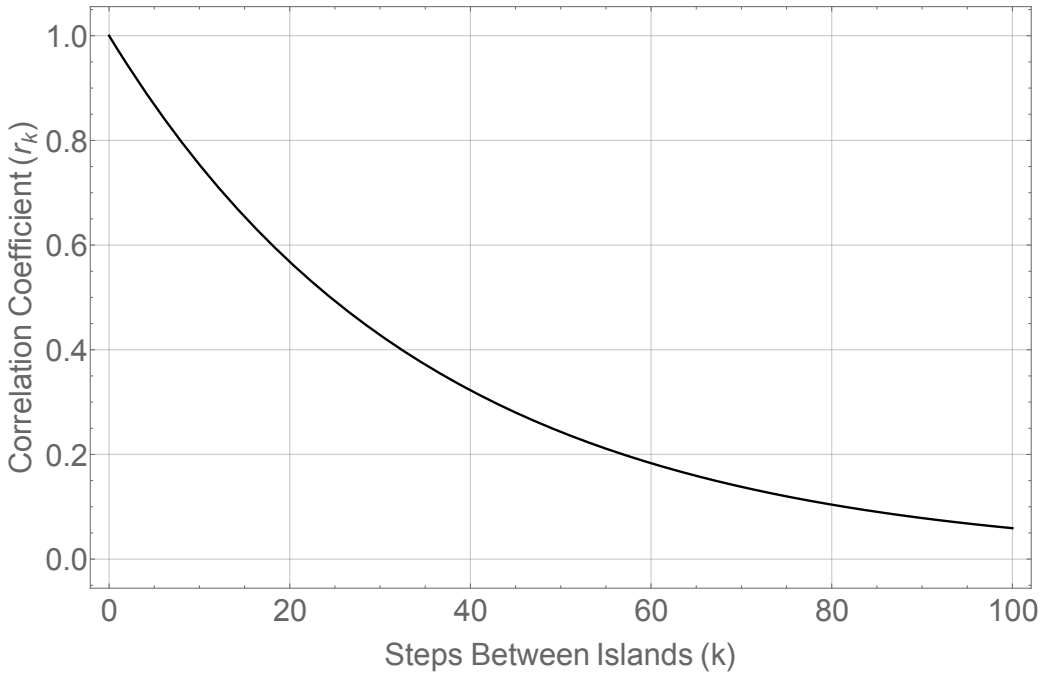


Figure 2.2: Kimura and Weiss's theoretical exponential decay of the correlation coefficient (r_k) with increasing physical distance (k) in the one-dimensional stepping-stone model. The rates of migration are $m_1 = 0.1$, $m_\infty = 4 \times 10^{-5}$.

This theoretical exponential decay has become an expected result in the field. However, there are a few factors that should be taken into consideration. First, Kimura and Weiss assume that the variance in allele frequency does not change from generation to generation. In natural populations, this variance may change. Second, they do not provide an estimate of how long it takes for a population to converge to this steady state— this will be explored further in Section 2.3. Third, and perhaps most noticeable, is the fact that Kimura and Weiss use an infinite system of islands, which is not a physically possible scenario.

This prompted us to investigate whether or not Kimura and Weiss's results still hold true in finite cases. We used a capped linear stepping-stone structure, which is exactly what its name brings to mind: a line of colonies without migration or communication between the ends. This model was explored in the 1970s by Maruyama, who used a matrix system to investigate its behavior.

2.2 Matrix Analysis

Maruyama [11] analyzed the finite linear system of islands with a variety of different boundary conditions (describing migration to and from the end islands). Rather than using the correlation coefficient r_k from Kimura and Weiss, Maruyama chose to use the covariance between pairs of islands. If p_i is the allele frequency on the ith island and $\delta_i = p_i - \bar{p}$ is the deviation of the ith frequency from its expectation $\bar{p}$, then the covariance between island i and island j is $c_{ij} = \text{cov}(\delta_i, \delta_j)$.

Naturally we would like to see how these quantities change over the course of generations. Maruyama developed recurrence relations for the allele frequencies, δ_i's, and covariances to establish the system's evolution in matrix form. The case we considered was a system of n islands with N individuals apiece, where the end islands are what Maruyama defined as "absorbing boundaries" (Figure 2.3). The hypothetical colony shown represents the mixture of the full population, and is the source of long-range migration in the model. The allele frequency on this island is assumed to always remain at $\bar{p}$, the expectation of all frequencies in the total population. Note that the rate of one-step migration is m (this is balanced, with $\frac{1}{2}m$ to each of the right and left islands) and the rate of long-range migration is m_∞.

Maruyama was able to express the evolution of the covariances c_{ij} by storing them in a matrix $\mathbf{Q}$ and determining that in the next generation,

$$\mathbf{Q}' = (\alpha\mathbf{I} + \beta\mathbf{S})\mathbf{Q}(\alpha\mathbf{I} + \beta\mathbf{S}) + \mathbf{P}/2N \tag{2.3}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{P}$ is a diagonal matrix with entries $p_{ii} = \bar{p}(1-\bar{p}) - c_{ii}$, and the matrix $\mathbf{S}$ is uniquely determined by the type of migration in the population: when individuals can migrate one step to the right or left, this matrix has entries on the subdiagonal and superdiagonal. Maruyama expresses this type of migration using the

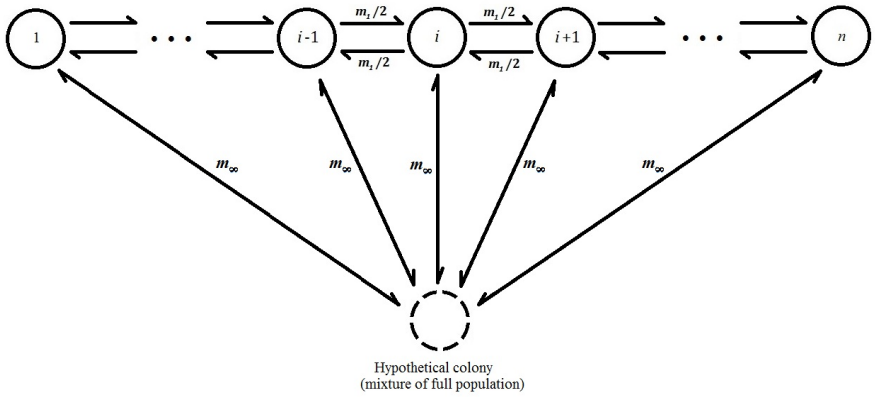


Figure 2.3: A finite system of islands with absorbing boundaries

$n \times n$ matrices

$$\mathbf{U} = \begin{bmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & \ddots \\ \vdots & & & \end{bmatrix} \quad \text{and} \quad \mathbf{U^T} = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \\ 0 & 1 & 0 & \\ \vdots & & \ddots & \end{bmatrix} \quad \Rightarrow \quad \mathbf{S} = \mathbf{U}+\mathbf{U^T} = \begin{bmatrix} 0 & 1 & 0 & \cdots \\ 1 & 0 & 1 & \\ 0 & 1 & 0 & \ddots \\ \vdots & & \ddots & \end{bmatrix}. \tag{2.4}$$

At equilibrium, Equation 2.3 becomes

$$\mathbf{Q} = (\alpha\mathbf{I} + \beta\mathbf{S})\mathbf{Q}(\alpha\mathbf{I} + \beta\mathbf{S}) + \mathbf{P}/2N, \tag{2.5}$$

which can be rewritten as a linear transformation from the set of all $n \times n$ matrices onto itself, where $L[\mathbf{Q}] = \mathbf{P}/2N$. For any matrix A, the linear operator would be $L[A] = A - (\alpha\mathbf{I} + \beta\mathbf{S})A(\alpha\mathbf{I} + \beta\mathbf{S})$.

It turns out that $\mathbf{Q}$ can be written as a linear combination of the eigenfunctions of $L[\ ]$, which in turn depend on the eigenvectors of the matrix $\mathbf{S}$. To determine the eigensystem of $\mathbf{S}$ (method from [6]), begin by setting up the eigenvector equation for $\mathbf{S}$ with eigenvectors $\mathbf{v}$ and eigenvalues λ:

$$\mathbf{Sv} = \lambda\mathbf{v} \Rightarrow (\mathbf{S} - \lambda I)\mathbf{v} = \mathbf{0}.$$

For reasons that will become apparent, we choose to write $\lambda = 2c$. Then we have

$$\Rightarrow \begin{bmatrix} -2c & 1 & 0 & \cdots & 0 \\ 1 & -2c & 1 & & \\ 0 & 1 & -2c & 1 & \\ \vdots & & & \ddots & \\ 0 & & & 1 & -2c \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

which, if we include two invented boundary variables $v_0 = v_{n+1} = 0$, produces

$$\begin{bmatrix} v_0 - 2cv_1 + v_2 \\ v_1 - 2cv_2 + v_3 \\ \vdots \\ v_{k-1} - 2cv_k + v_{k+1} \\ \vdots \\ v_{n-2} - 2cv_{n-1} + v_n \\ v_{n-1} - 2cv_n + v_{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

These equations are all of the form

$$v_{k-1} - 2cv_k + v_{k+1} = 0, \tag{2.6}$$

which is a second order linear difference equation with constant coefficients. This can be solved by using the ansatz $v_k = r^k$: the characteristic polynomial is $1 - 2cr + r^2 = 0$, with roots $r_\pm = c \pm \sqrt{c^2 - 1}$. Recall that we do not yet know the value of c (we are looking for it so that we may find the eigenvalues $\lambda = 2c$). There are two cases we must address for the value of c.

Case 1: $c = \pm 1$. This means that $r_+ = r_- = c$. Since the roots are not distinct, the general solution to the difference equation (2.6) is of the form $v_k = Ar^k + Bkr^k = (A + Bk)r^k$, or

$$v_k = (A + Bk)c^k$$

where A and B are constants. Using the boundary condition $v_0 = 0$, we find that $A = 0$. The other boundary condition $v_{n+1} = 0$ produces $0 = B(n+1)c^{n+1}$, which is only satisfied when $B = 0$. Therefore, in this case, we obtain only the trivial solution $v_k = 0$.

Case 2: $c \neq \pm 1$. Here, we have distinct values for r_+ and r_-, so the general solution to (2.6) is of the form $v_k = Ar_+^k + Br_-^k$. However, $r_- = c - \sqrt{c^2 - 1} = 1/(c + \sqrt{c^2 - 1}) = 1/r_+$, so by letting $r_+ = r$ we may rewrite the solution as

$$v_k = Ar^k + Br^{-k}.$$

Plugging in the boundary condition $v_0 = 0$, we obtain $A + B = 0$, so then the solution is $v_k = A(r^k - r^{-k})$. The other boundary condition, $v_{n+1} = 0$, implies that $A(r^{n+1} - r^{-(n+1)}) = 0$. $A = 0$ is the trivial solution, so we must have $r^{n+1} - r^{-(n+1)} = 0 \Rightarrow r^{2(n+1)} = 1$. In order for this to hold, $|r| = 1$.

Using the fact that $|r| = 1$, we can write $r = e^{i\theta}$, so $r^{2(n+1)} = 1 \Rightarrow e^{2i(n+1)\theta} = 1$. From this, $2(n+1)\theta = 2k\pi \Rightarrow \theta = \frac{k\pi}{n+1}$ for some $1 \leq k \leq n$. Since $r = c + \sqrt{c^2 - 1}$, we can think of c as $\cos\theta$. This provides what we need to find the eigenvalues, since we let $\lambda = 2c$:

$$\lambda_k = 2\cos\theta = 2\cos\frac{k\pi}{n+1} \quad \text{for } \ k = 1, 2, ..., n. \tag{2.7}$$

This differs from Maruyama's original calculation by a factor of 2, but that constant does

not appear to cause a difference in the results. The behavior of the system depends on the migration, and therefore on these eigenvalues. To investigate the convergence of the system, we can look at the ratio of the largest and second-largest eigenvalues [14].

Since $k = 1, 2, \ldots, n$, the eigenvalue with the largest magnitude is $|\lambda_1| = |\lambda_n| = 2|\cos\frac{\pi}{n+1}|$. The next largest is $|\lambda_2| = |\lambda_{n-1}| = 2|\cos\frac{2\pi}{n+1}|$, and the ratio of the largest to second-largest eigenvalues is

$$\left|\frac{\lambda_2}{\lambda_1}\right| = \left|\frac{\cos\frac{2\pi}{n+1}}{\cos\frac{\pi}{n+1}}\right|. \tag{2.8}$$

From this, we obtain an interesting result. This ratio will produce a larger number for larger values of n— for example, when $n = 5$, $|\lambda_2/\lambda_1| = 1/\sqrt{3} \approx 0.577350$ and when $n = 1000$, $\lambda_2/\lambda_1 \approx 0.999985$. Since the ratio of the first two largest eigenvalues reflects the convergence rate, this means that population systems with fewer islands will converge to their steady state faster than population systems with a greater number of islands.

In addition to the eigenvalues of the migration matrix S, the eigenvalues of the linear operator $L[\,]$ (which describes the full dynamics of the population system) also contribute to a useful result. The eigenvalues of $L[\,]$ are given by

$$\xi_{kl} = 1 - (1 - m_\infty)^2 \left[1 - m\left(1 - \cos\frac{\pi k}{n+1}\right)\right]\left[1 - m\left(1 - \cos\frac{\pi l}{n+1}\right)\right]. \tag{2.9}$$

Maruyama [11] states that the long-range migration term m_∞ is necessary to stabilize the system, and that "without this term there is no meaningful stationary correlation of the gene frequencies between colonies," but he provides no justification for this claim. It is here that we are finally able to verify the statement. Generally speaking, if all eigenvalues of a system are positive, then the system is unstable at equilibrium. We consider the case now where $m_\infty = 0$. Then the eigenvalues of the linear operator $L[\,]$ are

$$\begin{aligned}
\xi_{kl} &= 1 - \left[1 - m\left(1 - \cos\frac{\pi k}{n+1}\right)\right]\left[1 - m\left(1 - \cos\frac{\pi l}{n+1}\right)\right] \\
&= 1 - \left[1 - m\left(1 - \cos\frac{\pi l}{n+1}\right) - m\left(1 - \cos\frac{\pi k}{n+1}\right)\right. \\
&\qquad \left. + m^2\left(1 - \cos\frac{\pi k}{n+1}\right)\left(1 - \cos\frac{\pi l}{n+1}\right)\right] \\
&= m\left(1 - \cos\frac{\pi l}{n+1}\right) + m\left(1 - \cos\frac{\pi k}{n+1}\right) - m^2\left(1 - \cos\frac{\pi k}{n+1}\right)\left(1 - \cos\frac{\pi l}{n+1}\right) \\
&= m\left(2 - \cos\frac{\pi l}{n+1} - \cos\frac{\pi k}{n+1}\right) - m^2\left(1 - \cos\frac{\pi k}{n+1}\right)\left(1 - \cos\frac{\pi l}{n+1}\right).
\end{aligned}$$

The variable m represents the one-step migration rate, so we know that $0 < m < 1$ and thus $m^2 < m$. We also prove the following result, beginning with the fact that $-1 \leq$

$\cos\frac{\pi k}{n+1}, \cos\frac{\pi l}{n+1} \leq 1$:

$$1 \geq \left(\cos\frac{\pi l}{n+1}\right)\left(\cos\frac{\pi k}{n+1}\right)$$

$$2 - \cos\frac{\pi l}{n+1} - \cos\frac{\pi k}{n+1} \geq 1 - \cos\frac{\pi l}{n+1} - \cos\frac{\pi k}{n+1} + \left(\cos\frac{\pi l}{n+1}\right)\left(\cos\frac{\pi k}{n+1}\right)$$

$$2 - \cos\frac{\pi l}{n+1} - \cos\frac{\pi k}{n+1} \geq \left(1 - \cos\frac{\pi k}{n+1}\right)\left(1 - \cos\frac{\pi l}{n+1}\right).$$

Since $m > m^2$ and $2 - \cos\frac{\pi l}{n+1} - \cos\frac{\pi k}{n+1} \geq \left(1 - \cos\frac{\pi k}{n+1}\right)\left(1 - \cos\frac{\pi l}{n+1}\right)$, we conclude that $m\left(2 - \cos\frac{\pi l}{n+1} - \cos\frac{\pi k}{n+1}\right) > m^2\left(1 - \cos\frac{\pi k}{n+1}\right)\left(1 - \cos\frac{\pi l}{n+1}\right)$. Therefore, when $m_\infty = 0$, the eigenvalues $\xi_{kl} > 0$ for all k, l and the system is unstable. For this reason, we included a nonzero m_∞ term in all simulations.

2.3 Results

We wanted to see if the assumption of specific exponential decay at equilibrium (as predicted by Kimura and Weiss, Eq. 2.2) is realistic. Using the code described in Appendix A, we have simulated the evolution of a subdivided population over many generations, periodically investigating the shape of the correlation coefficient curve. We have chosen to adhere to Kimura and Weiss's original values for the migration rate parameters: $m_1 = 0.1$ and $m_\infty = 4 \times 10^{-5}$.

2.3.1 Time to Convergence

One of the main things we wanted to determine was how long it takes for a population to reach equilibrium. Figure 2.4 shows the evolution of the correlation coefficient r_5 (the correlation between islands that are 5 steps apart) as a reflection of the system's behavior. After approximately 15,000 generations, the correlation appears to have reached a steady state. We can also observe that there is a transient period of exponential increase during the first 4,000-6,000 generations: this is why Maruyama chose to disregard the data from the first 2,000 generations of his simulations [11].

Now that we have verified that the system is approaching some equilibrium state, we want to see if it is approaching the same state that was predicted by Kimura and Weiss. Figure 2.5 compares the simulated correlation coefficient curves to Kimura and Weiss's theoretical result (Eq. 2.2). The simulated curves show the average value of r_k (from 500 trials) corresponding to values of k from 0 to 100, and they reflect the state of the population at 1,000-generation time intervals. The initial correlation curve lies along the k-axis with a correlation of approximately zero, which is to be expected (there is not a relationship between p_i on different islands at first). Observe that the simulated result exhibits exponential decay, as predicted, and as time progresses the simulated values are approaching the theoretical values. The difference between consecutive curves decreases,

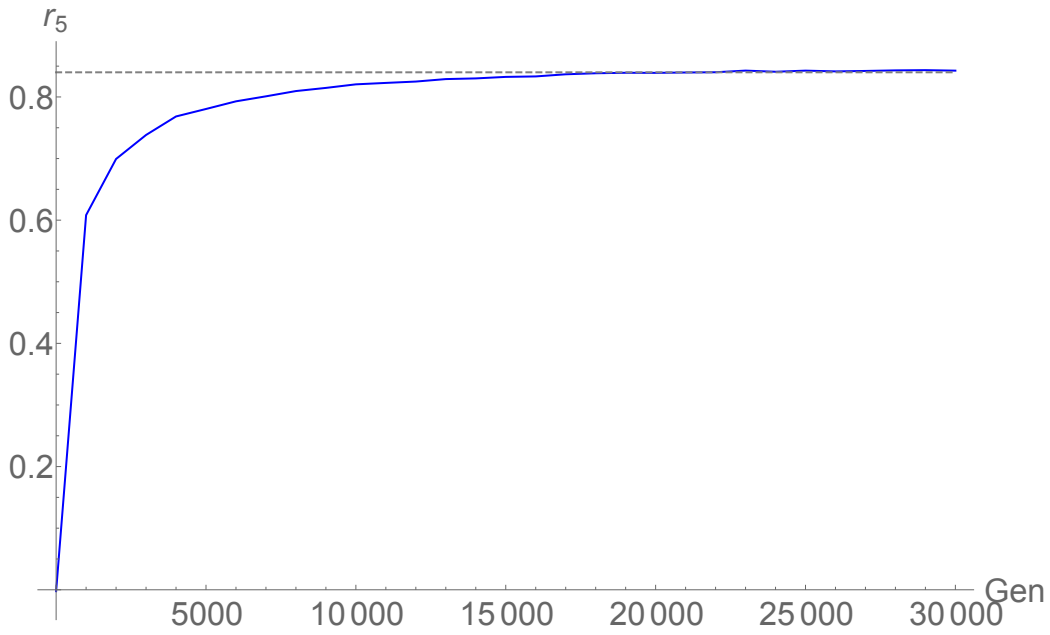


Figure 2.4: The correlation between pairs of islands that are 5 steps apart (r_5) appears to approach a steady state after approximately 15,000 generations

indicating that we are asymptotically approaching the equilibrium state. Our next objective is to track this approach by quantifying the difference between simulation and theory.

In Figure 2.6, we compare the simulated and theoretical results by tracking changes in the residual sum of squares (RSS) every 1,000 generations. The RSS is calculated by the following:

$$\text{RSS} = \sum_{k=0}^{100} (r_{k,\text{simulated}} - r_{k,\text{Kimura-Weiss}})^2 \tag{2.10}$$

After an initial period of rapid decrease, the RSS stabilizes at a value near 0.5. From the shape of the RSS curve, we can tell that the simulation is asymptotically approaching a steady-state solution. However, since the RSS stabilizes near a nonzero value, the steady-state solution may not be exactly the result predicted by Kimura and Weiss.

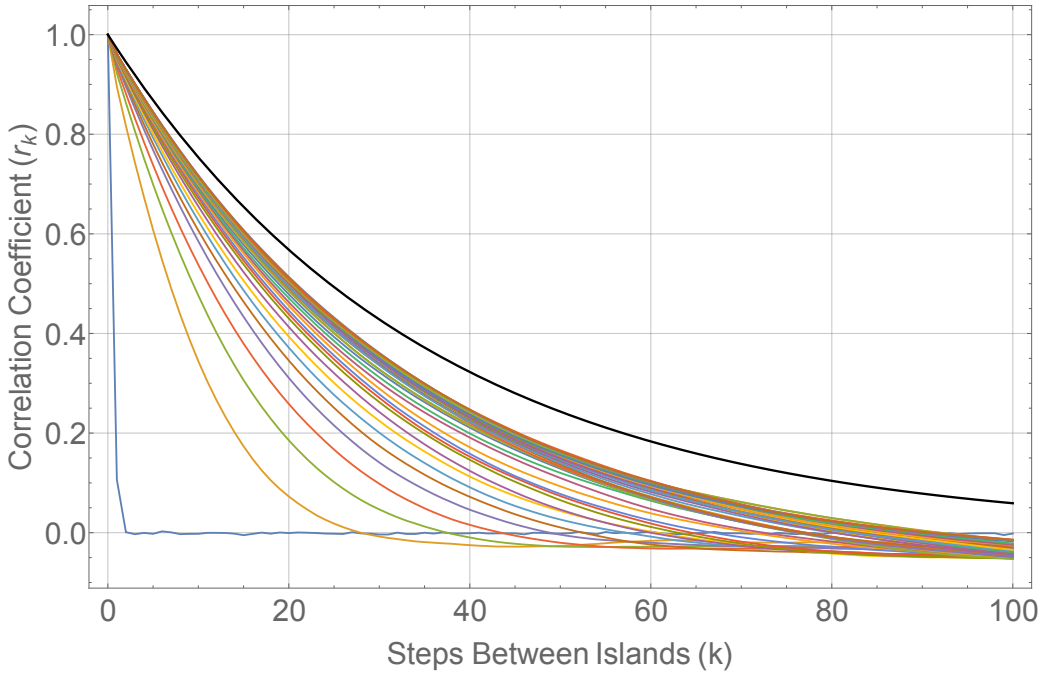


Figure 2.5: Average correlation coefficient values for islands k steps apart (linear population structure, $n = 1000$), sampled 50 times over 50,000 generations, asymptotically approach Kimura and Weiss's theoretical result (shown in black).

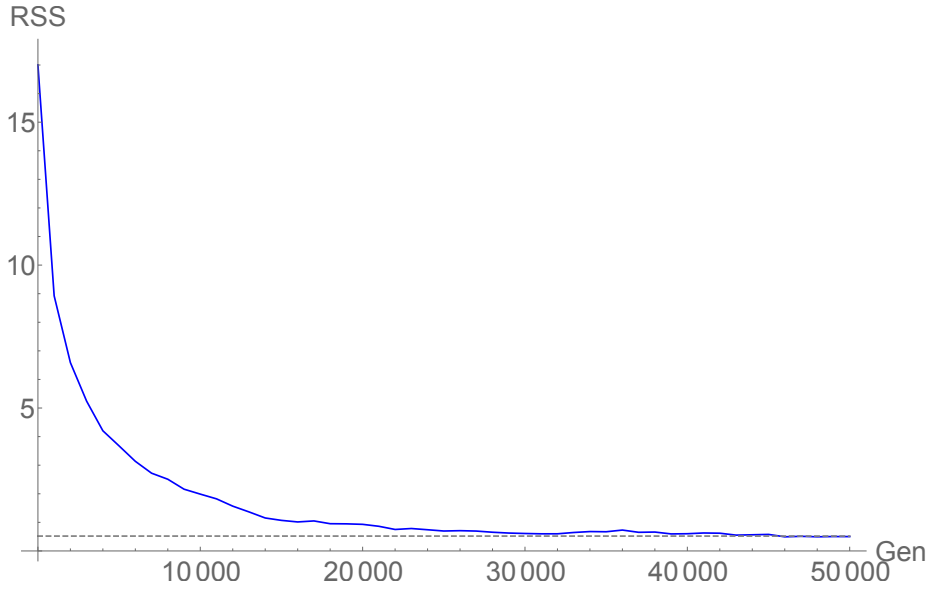


Figure 2.6: Over the course of generations, the residual sum of squares measures the difference between the simulated data (linear model) and the theoretical result by Kimura and Weiss. The RSS value stabilizes around 0.52.

One assumption made by Kimura and Weiss is that the variance of allele frequencies within the population is unchanging. As we show in Figure 2.7, this is not necessarily the case at all points in time. The variance undergoes a period of exponential increase at first, which may have been disregarded in previous results (particularly by Maruyama) since it is transient behavior. By the time we appear to have reached a steady state in correlation coefficient values (approximately 15,000-20,000 generations), the variance is still experiencing gradual changes but has generally converged to a value near 0.10.

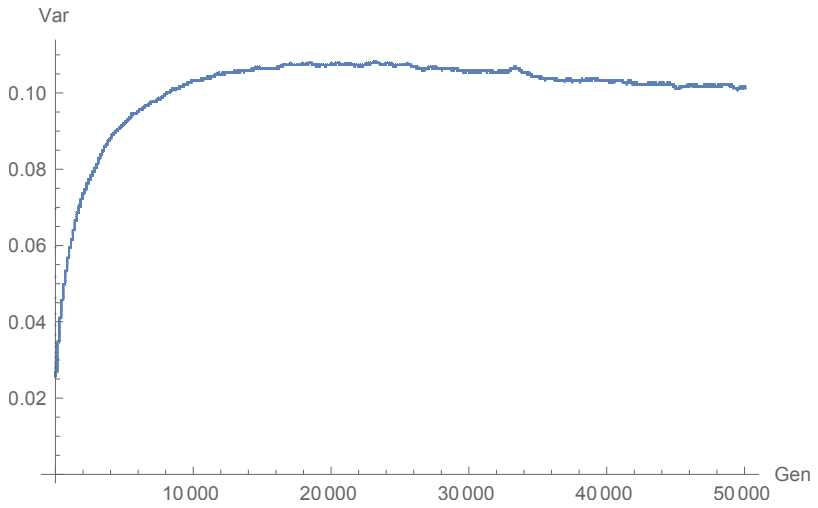


Figure 2.7: The evolution of the variance in the linear model, evolving over 50,000 generations.

2.3.2 Number of Islands

Now that we see that the variance is not the likely cause of the difference between the simulated and theoretical steady-state solutions, we turn to another of Kimura and Weiss's assumptions: the fact that their result is for an infinite number of islands. This leads us to expect that using a smaller number of islands in the simulation will increase the difference between the simulated and theoretical results.

In Figure 2.8, we see that using $n = 500$ islands rather than $n = 1000$ does indeed produce a simulated result that is farther from Kimura and Weiss's at the 20,000-generation mark. Soon after this point in time, though, we encounter an interesting issue that cannot be observed graphically.

Recall that the speed of convergence of the system depends upon the number of islands, where systems with smaller n have faster convergence (see Equation 2.8). This holds true, and the system with 500 islands does converge faster. However, the steady-state result is not Kimura and Weiss's decay curve— instead, we obtain is what is known as allele fixation. In fixation, the allele of interest either permeates the entire population (so every island has an allele frequency of $p_i = 1$) or it is eliminated completely (so every island has $p_i = 0$).

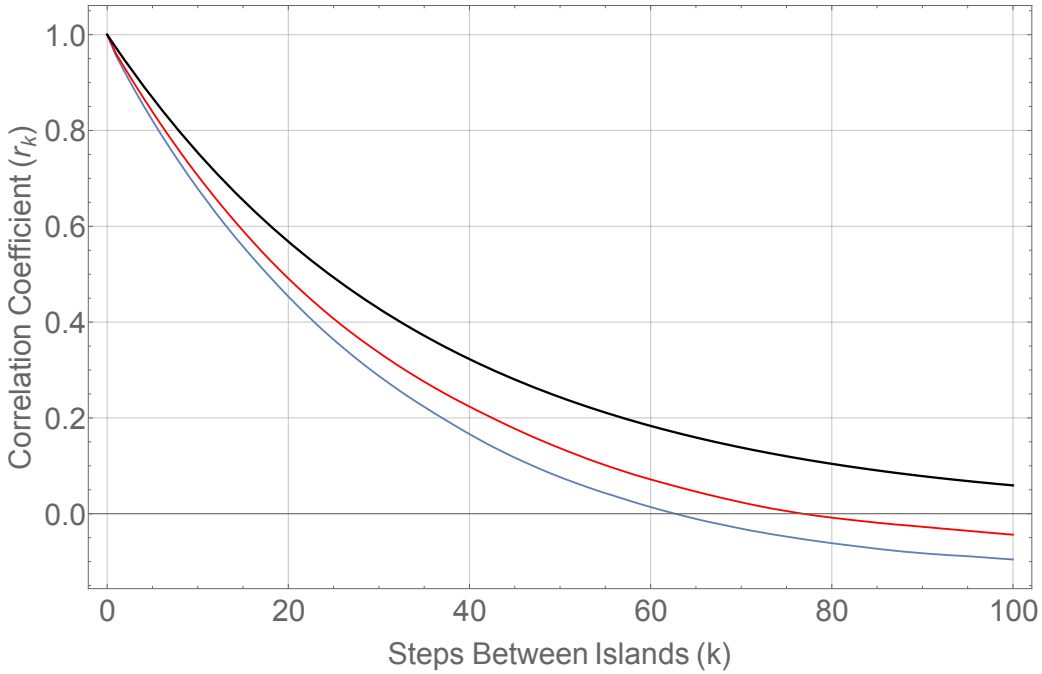


Figure 2.8: At 20,000 generations: the average correlation coefficient values for the linear model with $n = 500$ islands (blue) or $n = 1000$ islands (red), compared to Kimura and Weiss's result (black).

Fixation does not occur every time the simulation is run, due to the randomness involved in the simulation process, but it does occur significantly more often in systems with fewer islands. When $n = 500$, 17 out of 500 runs ended in fixation, compared to 0 out of 500 runs when $n = 1000$.

Let us now look at a system that contains $n = 3000$ islands. We see in Figure 2.9 that the system is definitely approaching the theoretical result predicted by Kimura and Weiss (Eq. 2.2), but convergence is taking a long time. None of the simulations encountered fixation, even after 100,000 generations. Figure 2.10 provides a snapshot similar to Figure 2.8 by comparing models with $n = 500, 1000$, and 3000 at the 20,000-generation mark. We can see that even at that relatively early point in time, the simulation with the greatest number of islands comes closest to matching Kimura and Weiss's result.

2.4 Summary

To generalize the results seen in this section, we have found that systems with a greater number of islands produce a simulated result that is closer to the theoretical result postulated by Kimura and Weiss (Eq. 2.2). However, increasing the number of islands also increases the time to convergence. Systems with a smaller number of islands may experience allele fixation, a steady state that is not as commonly seen in larger systems.

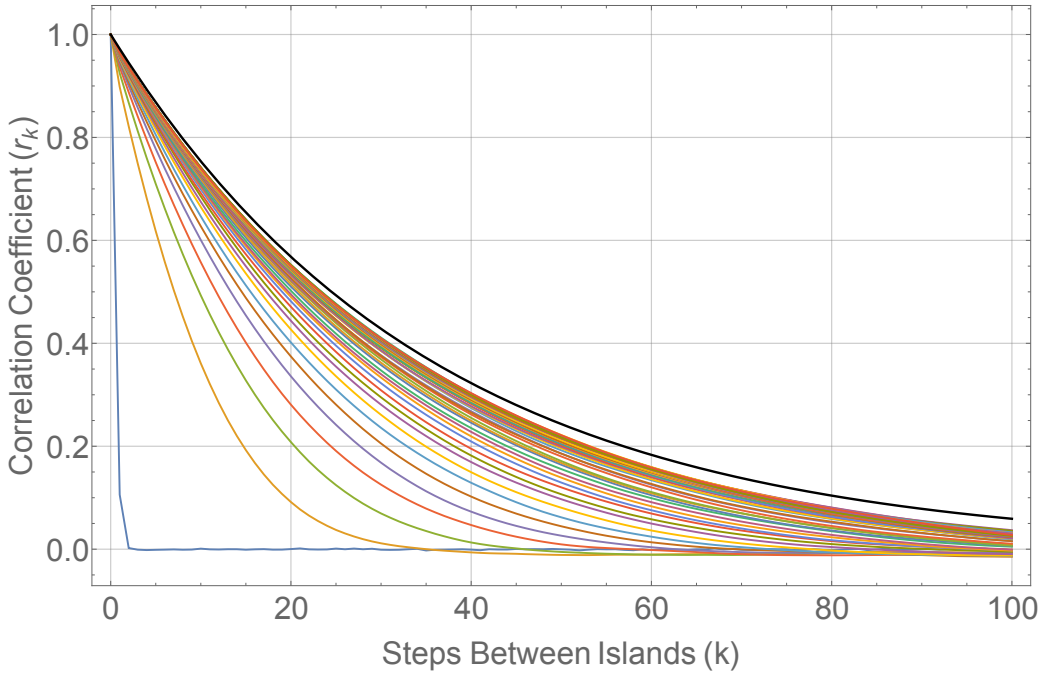


Figure 2.9: Average correlation coefficient values for islands k steps apart (linear population structure, $n = 3000$), sampled 100 times over 100,000 generations, asymptotically approach Kimura and Weiss's theoretical result (shown in black).

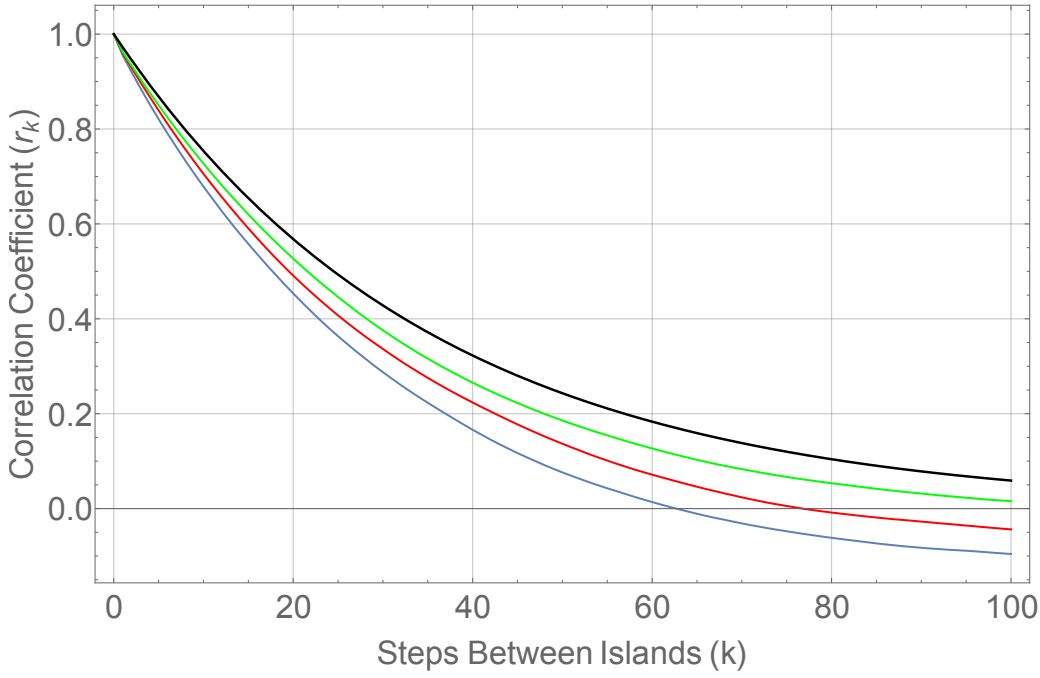


Figure 2.10: At 20,000 generations: the average correlation coefficient values for the linear model with $n = 500$ (blue),$n = 1000$ (red), or $n = 3000$ islands (green), compared to Kimura and Weiss's result (black).

Chapter 3

Circular Model

3.1 Introduction

Next, we considered a population with a different geographic structure. We applied the one-dimensional stepping stone model to a ring of islands (Figure 3.1) by allowing migration between what would have previously been the boundary islands of the linear structure. As in the linear model, one-step migration (rate $m_1 = 0.1$) and long-range migration (rate $m_\infty = 4 \times 10^{-5}$, from the mixture of the entire population) are considered.

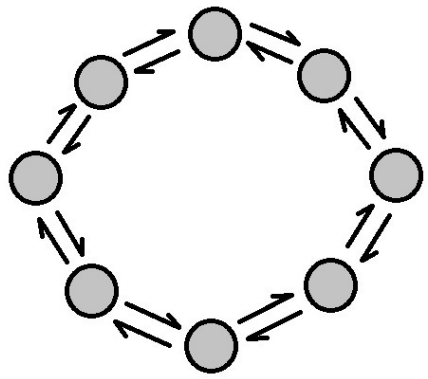


Figure 3.1: One-dimensional stepping stone model on a ring of islands.

Our goal, as before, was to simulate the population and determine the relationship between physical distance and genetic correlation, particularly once the system reaches equilibrium.

3.2 Matrix Analysis

Although Maruyama did explore the circular stepping-stone population structure, he did not use the same treatment as in the linear structure. He considered a case of unbalanced migration (where individuals were more likely to migrate in one direction than

the other) [10], and also investigated the general decrease of heterozygozity in the population as a whole [12]. For consistency, our analysis was based on Maruyama's treatment of the linear model seen in Section 2.2: we found the eigenvalues of the migration matrix to determine the system's rate of convergence.

Consider a circular system of n islands where individuals may migrate 1 step to the right or left. The migration in this case can be described by the following $n \times n$ matrix:

$$S_C = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 1 \\ 1 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

This matrix is circulant, which means that each of its rows is a cyclic permutation of the first row. The first row is denoted by the vector $\vec{c} = [c_0 \quad c_1 \quad c_2 \quad \cdots \quad c_{n-2} \quad c_{n-1}] = [0 \;\; 1 \;\; 0 \;\; \cdots \;\; 0 \;\; 1]$, so $c_j = 1$ if $j = 1$ or $j = n-1$, and $c_j = 0$ otherwise.

Circulant matrices have been well-studied and there are known forms for their eigenvalues and eigenvectors [3]. Specifically, the eigenvectors of a circulant matrix are given by

$$\lambda_j = c_0 + c_1\rho_j + c_2\rho_j^2 + \cdots + c_{n-1}\rho_j^{n-1}$$

where $j = 0, 1, \cdots, n-1$ and $\rho_j = \exp(-2\pi ij/n)$ are the nth roots of unity. In the case of M_C, we know that $c_j = 0$ except for $c_1 = c_{n-1} = 1$. Therefore the eigenvalues can be expressed as

$$\begin{aligned} \lambda_j &= \rho_j + \rho_j^{n-1} \\ &= e^{\frac{-2\pi ij}{n}} + e^{\frac{-2\pi ij(n-1)}{n}} \\ &= e^{\frac{-2\pi ij}{n}} + e^{-2\pi ij} e^{\frac{2\pi ij}{n}} \\ &= e^{\frac{-2\pi ij}{n}} + e^{\frac{2\pi ij}{n}} \\ &= \cos\left(\frac{-2\pi j}{n}\right) + i\sin\left(\frac{-2\pi j}{n}\right) + \cos\left(\frac{2\pi j}{n}\right) + i\sin\left(\frac{2\pi j}{n}\right) \\ &= 2\cos\left(\frac{2\pi j}{n}\right). \end{aligned}$$

Since $j = 0, 1, \cdots, n-1$, the eigenvalue with the largest magnitude is $|\lambda_0| = 2$. The next largest is $|\lambda_1| = |\lambda_{n-1}| = |2\cos\frac{2\pi}{n}|$, and the ratio of the largest to second-largest eigenvalues is

$$\left|\frac{\lambda_1}{\lambda_0}\right| = \left|\frac{2\cos\frac{2\pi}{n}}{2}\right| = \left|\cos\frac{2\pi}{n}\right|. \tag{3.1}$$

We see that the same result as in Section 2.3 holds: this ratio is larger when n is larger. When $n = 5$ islands, $\lambda_1/\lambda_0 \approx 0.309017$ and when $n = 1000$ islands, $\lambda_1/\lambda_0 \approx 0.999980$.

Since the value of the ratio approaches 1 as n increases, systems with a greater number of islands will once again take longer to converge. Note that while smaller systems will converge faster, they may still converge to a state of allele fixation (where an allele either permeates the entire population or is eliminated) more readily than a larger system.

Recall that in the linear system, the ratio of the two largest eigenvalues was 0.577350 when $n = 5$ and 0.999985 when $n = 1000$, as compared to 0.309017 and 0.999980 respectively in the circular system. So in both cases, the circular model is predicted to converge more rapidly than the linear model. To see if this pattern holds in general, let us compare the results of Eq. 3.1 to those of Eq. 2.8, the equivalent for the linear model.

Equation 2.8 states that the rate of convergence for the linear system is $\left|(\cos\frac{2\pi}{n+1})/(\cos\frac{\pi}{n+1})\right|$. Since $-1 \leq \cos\frac{\pi}{n+1} \leq 1 \Rightarrow \left|\cos\frac{\pi}{n+1}\right| \leq 1$, it follows that

$$\left|\cos\frac{2\pi}{n+1}\right| \leq \left|\frac{\cos\frac{2\pi}{n+1}}{\cos\frac{\pi}{n+1}}\right|$$

for all n. It can be seen that $\left|\cos\frac{2\pi}{n}\right| \leq \left|\cos\frac{2\pi}{n+1}\right|$ as long as $\frac{2\pi}{n} \leq \frac{\pi}{2} \Rightarrow n \geq 4$, so for all systems with 4 or more islands we will have

$$\left|\cos\frac{2\pi}{n}\right| \leq \left|\cos\frac{2\pi}{n+1}\right| \leq \left|\frac{\cos\frac{2\pi}{n+1}}{\cos\frac{\pi}{n+1}}\right|$$

$$\Rightarrow \left|\cos\frac{2\pi}{n}\right| \leq \left|\frac{\cos\frac{2\pi}{n+1}}{\cos\frac{\pi}{n+1}}\right|. \tag{3.2}$$

Therefore, between models with the name number of islands n ($n \geq 4$), the rate of convergence of the circular model will always be faster than that of the linear model. We will observe this result in the following section.

3.3 Results

3.3.1 Convergence

Figure 3.2 shows the progression of the simulated coefficient curve over time, and compares that to Kimura and Weiss's theoretical result (Eq. 2.2). We performed 500 runs of the simulation, each with a system of $n = 1000$ islands, and collected the correlation data every 1,000 generations for a total time of 50,000 generations. This result is highly similar to the one shown in Figure 2.5 for the linear model. Both exhibit exponential decay with increasing physical distance, and both are gradually approaching the result predicted by Kimura and Weiss.

Looking at the evolution of the value for the residual sum of squares (calculated by Eq. 2.10) over the course of 50,000 generations, we can see in Figure 3.3 that the RSS for

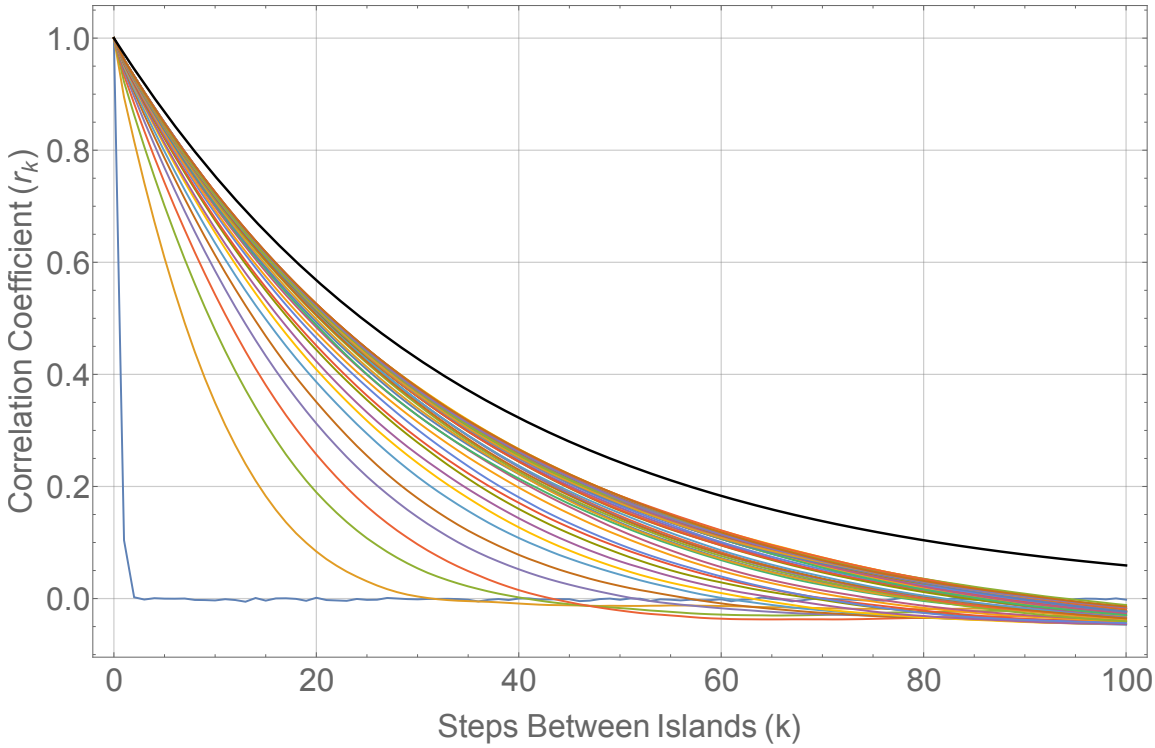


Figure 3.2: Average correlation coefficient values for islands k steps apart (circular population structure, $n = 1000$), sampled 50 times over 50,000 generations, asymptotically approach Kimura and Weiss's theoretical result (shown in black).

the circular system approaches a value near 0.37. Recall that the RSS for the linear system stabilized around 0.52. Therefore by this measure, the circular model was able to come closer to Kimura and Weiss's theoretical result by the 50,000th generation. This indicates that the circular model is converging more rapidly, which is consistent with the result found by comparing the convergence rates (ratio of the largest eigenvalues) for the two models, as seen in the previous section.

3.3.2 Comparison to Linear Model

Maruyama stated that "asymptotically the circular habitat can be considered as one-dimensional linear habitat" [11, p. 216], and indeed the two models do turn out to be very similar. In Figure 3.4, we see a snapshot of the correlation curves for the linear and circular models at the 50,000th generation, in comparison to Kimura and Weiss's correlation. Although both curves differ significantly from the theoretical result, they differ from one another by no more than 0.020, as seen in Figure 3.5.

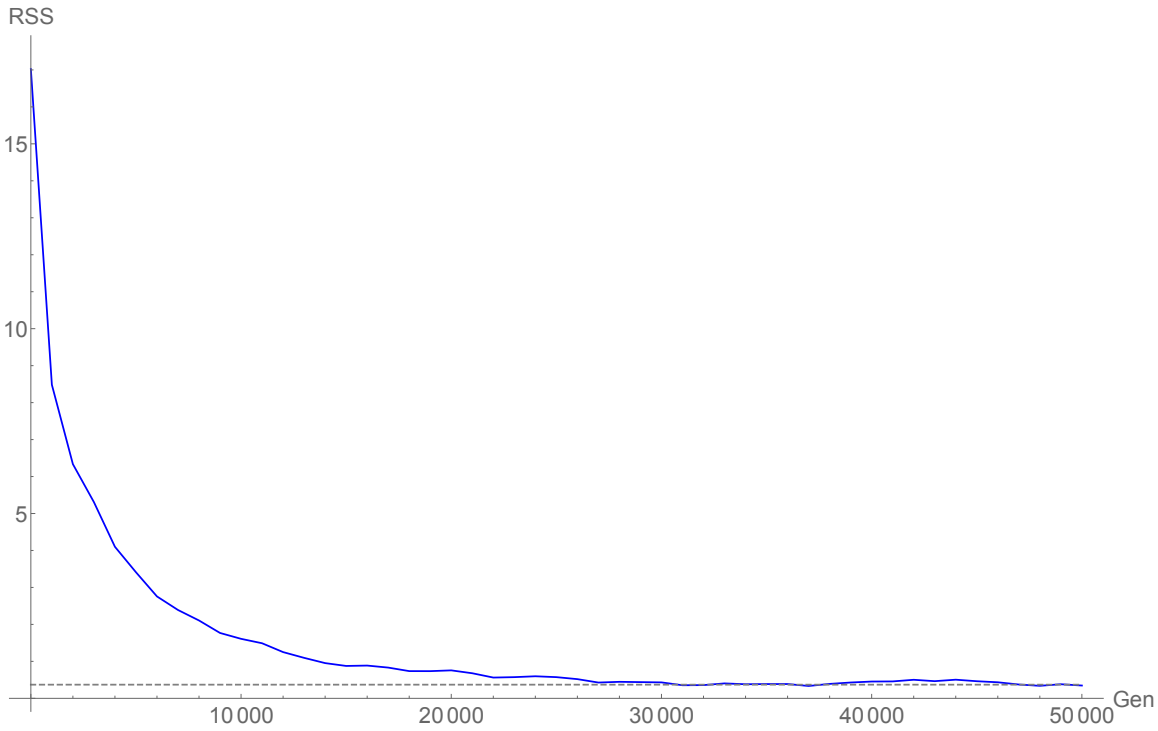


Figure 3.3: Residual sum of squares showing the difference between the simulated data (circular model, $n = 1000$) and the theoretical result by Kimura and Weiss. The RSS value stabilizes around 0.37.

3.4 Summary

Here we have seen that systems with a circular population structure also produce a simulated result that approaches Kimura and Wess's theoretical result, thereby verifying Maruyama's statement that the circular and linear models are asymptotically equivalent. Matrix analysis showed that again, systems with a greater number of islands will take longer to converge. When we compare the convergence rates of the circular and linear models with the same number of islands, the circular model converges more rapidly to the theoretical result.

One potential explanation for why the circular model provides a simulated result that more closely matches the theory is because the circular model overlaps on itself, thereby making it seem like the system has more islands than it actually does. Since the theoretical result is for an infinite number of islands, it makes sense that using a model that (even artificially) has "more" islands would provide better agreement.

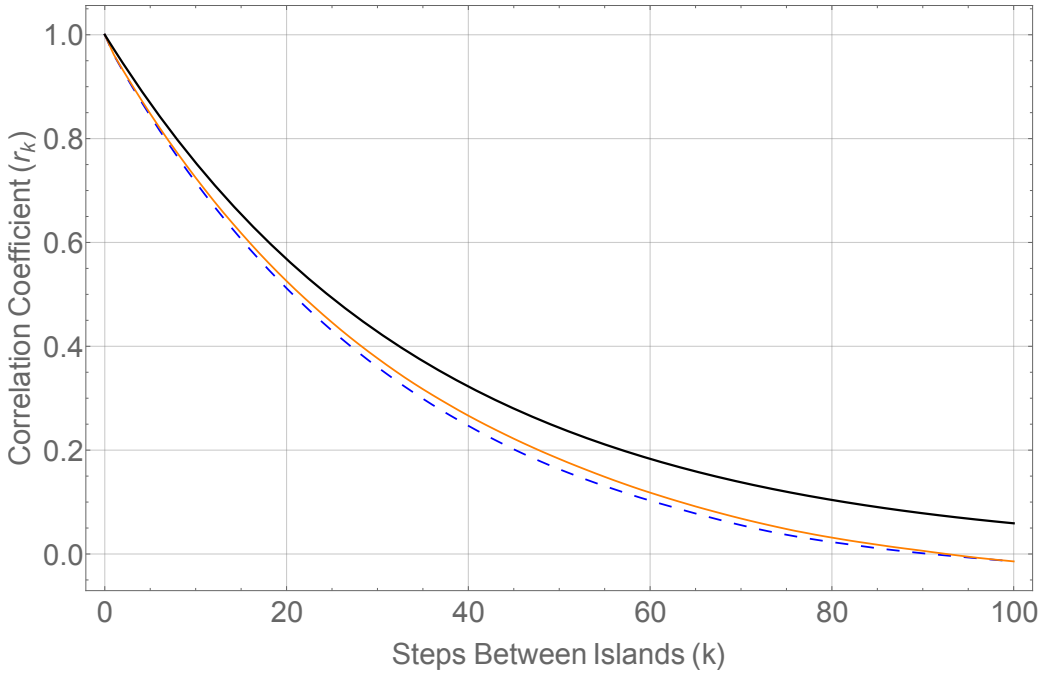


Figure 3.4: The linear model (blue, dashed curve) and the circular model (orange) at the 50,000 generation mark, as compared to Kimura and Weiss's result (black). Both simulated models have $n = 1000$ islands and identical migration rates.

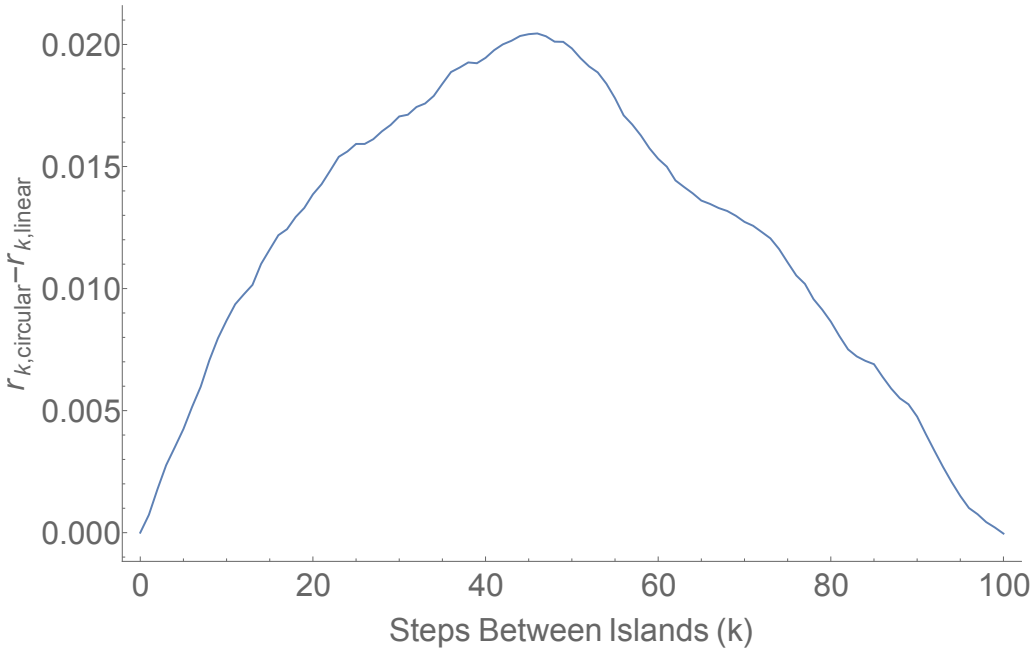


Figure 3.5: There is very little difference between the circular and linear models' average correlation coefficient for up to $k = 100$ steps at the 50,000th generation.

Chapter 4

Long-Range Migration Model

4.1 Introduction

In this chapter, we consider cases in which individuals can migrate further than one island away per generation. This is a more realistic scenario for many species. For example, it is common for marine species to spend part of their lives in a larval stage, during which individuals may be carried a variable distance away from their home colony [15]. The length of time spent in this stage is known as the pelagic duration, and it is thought to be a factor in predicting the genetic structure of a population due to its effect on the population's ability to disperse [1].

Kimura and Weiss developed a hypothesis for models with more general forms of migration [9], where they postulated that r_k had a form that was very similar to the case with one-step migration (Eq. 2.2). By "substituting for m_1 the variance of migration distance per generation", they obtained

$$r_k = e^{-\frac{\sqrt{2m_\infty}}{\sigma_m}k}, \tag{4.1}$$

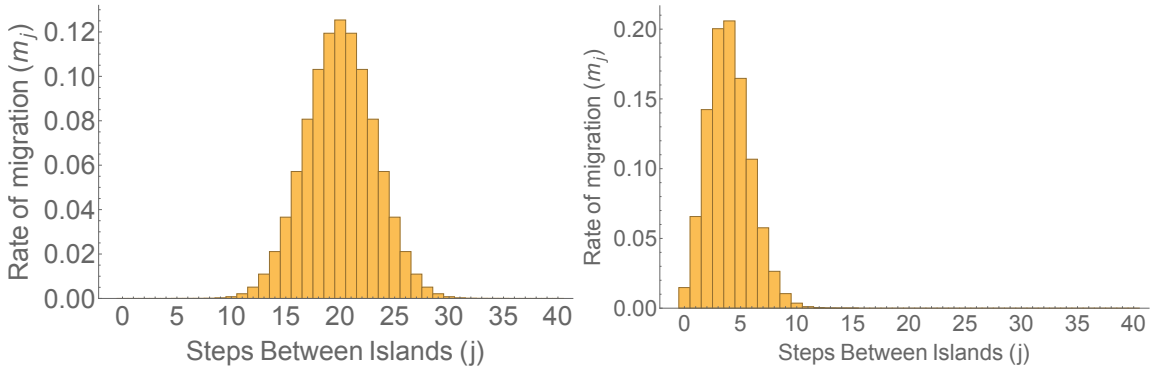
where σ_m is the variance of the migration distance, determined by

$$\sigma_m^2 = \sum_{j=1}^{\infty} j^2 m_j. \tag{4.2}$$

Plots of the theoretical result for selected migration-rate distributions are given in Figure 4.2. However, generally speaking, Kimura and Weiss's choice of $m_\infty = 4 \times 10^{-5}$ is a very small number, while σ_m is certainly larger than 1. Therefore by this estimation, r_k in the case of long-range migration will exhibit very slow exponential decay.

4.2 Numerical Simulation

First, we investigated a toy model in which individuals may only migrate 20 islands to the right or left each generation. This perfectly symmetric system is not biologically realistic, since organisms are not thought to count out the exact number of islands they may



(a) $m_j \sim B[40, 0.5]$ (b) $m_j \sim B[40, 0.1]$

Figure 4.1: Distributions used to simulate long-range migration. The number of steps an individual travels in a round of migration is selected from the desired distribution.

travel, but it does provide quantitatively interesting results.

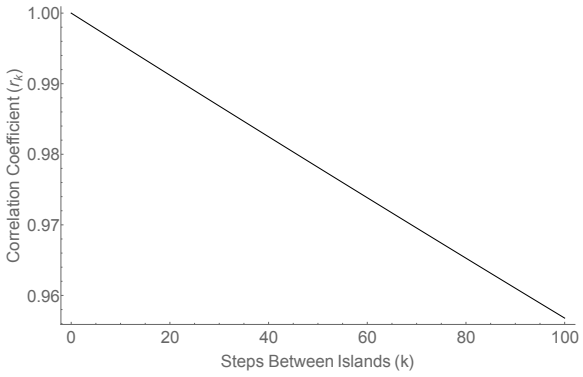
Next, we chose a more plausible migration pattern: one in which the number of steps an individual migrates (to the right or left) is chosen from a binomial distribution with a specified mean. Two binomial distributions (Figure 4.1) were selected for m_j, the rate of migration to islands j steps away: one was $m_j \sim B[40, 0.5]$, with mean $np = 20$ islands, and the other was $m_j \sim B[40, 0.1]$, with mean $np = 4$ islands. The plots showing the theoretical results for these distributions are shown in Figure 4.2. Note that Figures 4.2a and 4.2b have $k = 0, 1, ...100$, and only have a very small range in the r_k values. This shows how gradual the exponential decay is: out to $k = 100$, the curve decreases so little that it appears almost linear. In Figures 4.2c and 4.2d, we look at $k = 0, 1, ...1000$ so that we are finally able to see the exponential decay— in the case where $m_j \sim B[40, 0.1]$, at least. The decay rate in the case where $m_j \sim B[40, 0.5]$ is incredibly small: the theoretical equation describing this case with the given parameters is $e^{-0.000442k}$.

To incorporate the new style of migration, an option called `USE_BINOMIAL` was added to the standard C++ program included in Appendix A.2. If this option was specified to be true, the number of steps for an individual to move was selected from a choice of distribution in Figure 4.1. Otherwise, the migration was assumed to be "strict", where an individual can only move exactly a set number of islands— no more, no less.

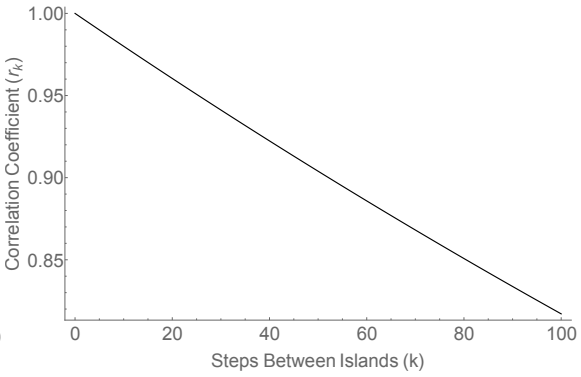
4.3 Results

4.3.1 Strict 20-Step Migration

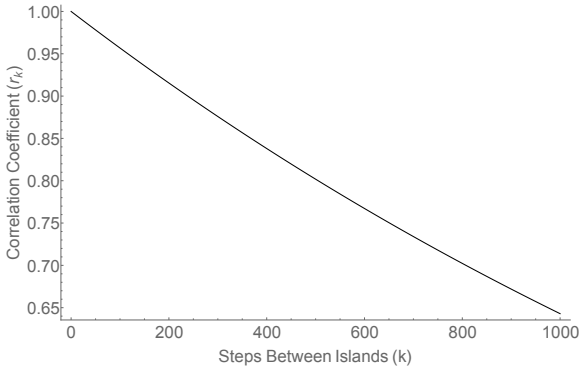
In this simplified model, individuals are only allowed to migrate to islands exactly 20 steps to the right or left of their current island. As seen in Figure 4.3, simulation provided a reasonable result for the genetic correlation: we can see that if the distance between islands is a multiple of 20 steps, those islands are much more highly correlated. It is logical that those islands have a high level of genetic similarity, because there is direct



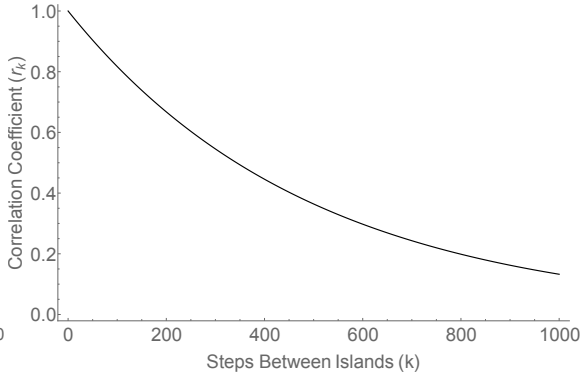
(a) $m_j \sim B[40, 0.5]$, out to $k = 100$ steps



(b) $m_j \sim B[40, 0.1]$, out to $k = 100$ steps



(c) $m_j \sim B[40, 0.5]$, out to $k = 1000$ steps



(d) $m_j \sim B[40, 0.1]$, out to $k = 1000$ steps

Figure 4.2: Kimura and Weiss's theoretical results for the decay of r_k in populations with longer-range migration (Eq. 4.1)

migration between them. While the snapshot in Figure 4.3 is taken at the 20,000th generation, there was not much variation in the correlation pattern before or after this point in time (the population began exhibiting the observed spikes in correlation as early as the 2,000th generation).

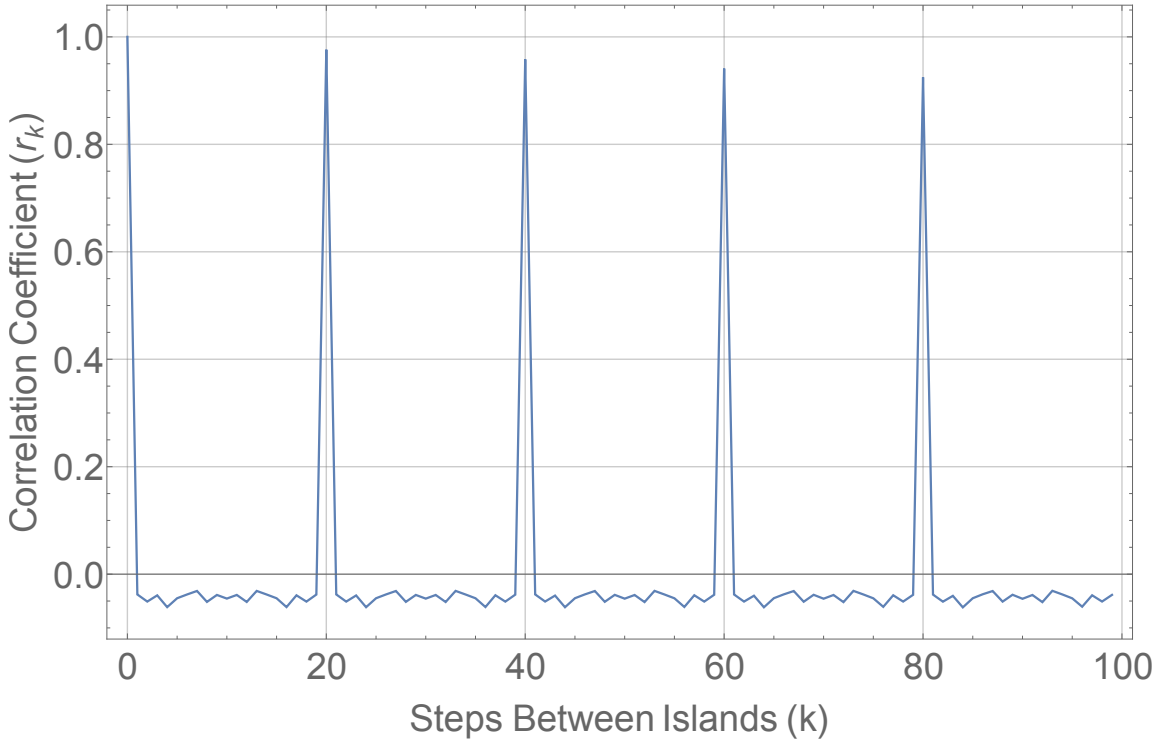


Figure 4.3: Average correlation coefficient values (at Generation 20,000) with strict 20-step migration.

4.3.2 Binomial Migration

In Figure 4.4, we see the result for the case with binomially-distributed migration with a mean of 20 steps. Observe that there is an increase in correlation when islands are approximately 20 steps apart, and faint ripples further out at 40 and 60 steps. This is in agreement with the pattern of migration, so the result seems intuitively logical.

However, when we compare the simulation to Kimura and Weiss's theoretical result (Eq. 4.1), the difference between the two is considerable. The simulated curves decidedly do not appear to approach the theoretical one, in stark contrast to the models seen in previous chapters. If we visually simplify the simulated curve to a line by excluding or flattening the areas with increases in correlation, we can observe that this resulting line would have approximately the same slope as the theoretical result. This may encourage us to suspect that perhaps the two methods have similar rates of decay, but that is where the likeness ends.

One of the most prominent differences between the theoretical and simulated results is the large discrepancy in their numerical values, even for small k. The simulated system has a steep initial drop in correlation, which indicates that the allele frequencies on even adjacent islands have little effect on each other. Again, considering the fact that migration in this system is not highly likely to send an individual to an adjacent island, this seems reasonable. Yet the theoretical result strongly disagrees. This leads us to suspect that perhaps this particular result of Kimura and Weiss's is not designed to reflect the type of migration used in the simulation.

Indeed, it turns out that Kimura and Weiss designed Equation 4.1 to be used when the long-range migration is "sufficiently weak," a distinction that they mention in a later paper [17]. The migration in this model must not meet those criteria, therefore a different theoretical result should be used.

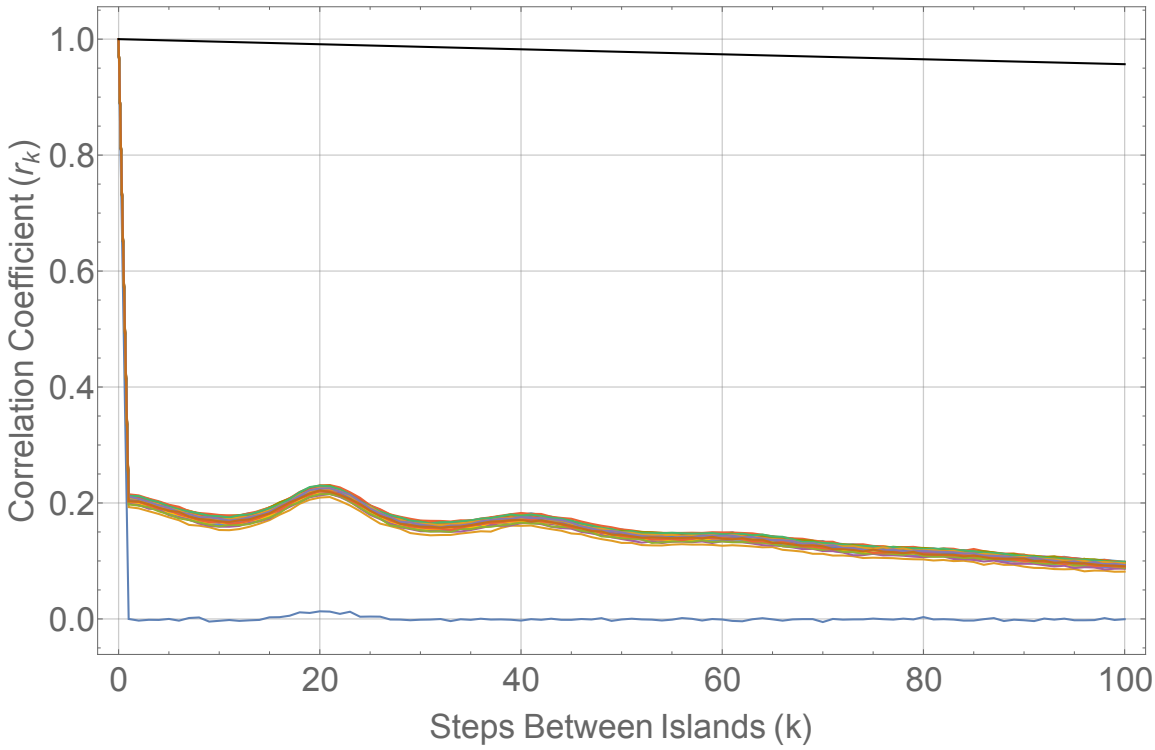


Figure 4.4: Average correlation coefficient values for a system with migration that is binomially distributed ($m_j \sim B[40, 0.5]$) with a mean of 20 islands. Kimura and Weiss's theoretical result is shown in black.

Although the mean number of steps traveled is closer to 1, the simulation with migration rates distributed by $m_j \sim B[40, 0.1]$ (Figure 4.5) still does not approach the curve given by Equation 4.1— this migration must not be sufficiently weak either.

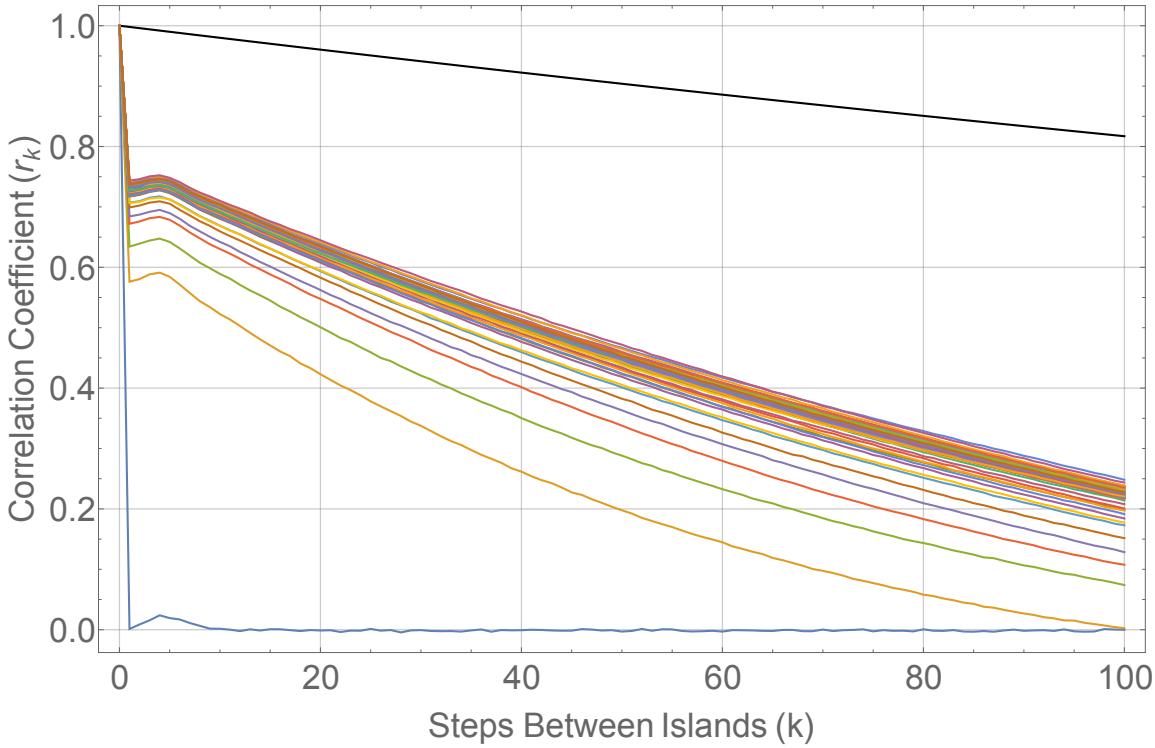


Figure 4.5: Average correlation coefficient values for a system with migration that is binomially distributed ($m_j \sim B[40, 0.1]$) with a mean of 4 islands. Kimura and Weiss's theoretical result is shown in black.

4.4 Summary

Simulations where migration rates are distributed according to $m_j \sim B[40, 0.5]$ and $m_j \sim B[40, 0.1]$ provide results that are intuitively logical, but do not converge to the theoretical result in Equation 4.1. It turns out that neither model has the sufficiently weak levels of long-range migration that allow for use of this equation. Further study is required to find the theoretical result to which these models should be compared.

Chapter 5

Conclusion and Future Work

We have successfully created a program to simulate the generational processes of migration and reproduction in the one-dimensional stepping-stone model of population structure. The program calculates r_k, a measure of genetic correlation between islands that are k steps apart. Through the simulation results, we were able to validate the general exponential decay of genetic correlation with increasing physical distance. In the cases with one-step migration, we were able to show that the simulation result asymptotically approaches a theoretical result hypothesized by Kimura and Weiss in the 1960s (Eq. 2.2). Both circular and linear population structures will converge to this result over a long period of time, provided that the number of islands is large enough (on the order of 10^3). In some systems, particularly ones with a smaller number of islands, simulation occasionally ended in fixation of an allele (where either $p_i = 0$ or $p_i = 1$ for all islands i) and the theoretical result was not attained.

Simulated results were consistent with matrix analysis of the system: population systems with a larger number of islands take a longer time to converge, but they more closely match the theoretical result (which was established for a system with an infinite number of islands). When given the same number of islands, the circular model converges slightly faster. This is observed not only in the simulation results, but also in the value of the residual sum of squares and in the convergence rates given by the ratio of the largest eigenvalues of the respective migration matrices.

In the case of long-range migration, the simulated results were found to be intuitively logical but showed no signs of converging to the anticipated theoretical result. Upon further inspection, the long-range migration patterns chosen may not be "sufficiently weak" to allow comparison to that particular equation.

These results are of interest in relation to fieldwork studies, since the theory by Kimura and Weiss is not based upon a physically possible population. Real populations only contain a finite number of islands, follow a diverse variety of migration patterns, and results are not considered in the asymptotic sense. Therefore the long convergence times seen in the simulation may not be biologically realistic— biologists in the field may want to consider how long the "generation" time is for a given organism, what migration pattern(s) are involved, and how long a particular population has been evolving when examining their

data.

This problem is a good candidate for further study because there are many remaining opportunities for analysis and application. In the theoretical realm, one could develop a time-dependent result for the finite models. It would be desirable to see if the simulation accurately converged to such a model at various points in time, rather than solely considering the asymptotic approach required by Kimura and Weiss. It would also be useful to resolve the differences between the theoretical and simulated results for the long-range migration case.

The simulation is open to a variety of modifications. First, the migration rates may be described by any other choice of distribution, which allows for customization to fit a given population. Also, a reproductive pattern may be implemented, which could be used to reflect certain reproductive events such as "blooms": periods of unusually rapid reproduction often seen in marine species [13]. Finally, other evolutionary processes such as mutation or selection may be incorporated.

There is also the potential to apply the outcomes of this project to field data on jellyfish populations collected by Professor Michael Dawson's lab at UC Merced. The simulation could be run with the parameters for jellyfish migration and reproduction to determine what genetic correlation patterns could be expected for a given population. Overall, we have designed the simulation to be able to accommodate a wide range of population characteristics, so that it may be used in a variety of future applications.

Bibliography

[1] Michael N Dawson, Cynthia G Hays, Richard K Grosberg, and Peter T Raimondi. Dispersal potential and population genetic structure in the marine intertidal of the eastern north pacific. Ecological Monographs, 84(3):435–456, 2014.

[2] Douglas J. Futuyma. Evolutionary Biology. Sinauer Associates, 3rd edition, 1998.

[3] Robert M Gray. Toeplitz and circulant matrices: A review. now Publishers Inc., 2006.

[4] Richard Halliburton. Introduction to Population Genetics. Pearson/Prentice Hall, 2004.

[5] Godfrey H. Hardy et al. Mendelian proportions in a mixed population. Science, 28(706):49–50, 1908.

[6] Jerry L. Kazdan. A tridiagonal matrix, 2010. [Online; accessed 3-February-2015].

[7] M Kimura. "stepping-stone" model of population. Annual Report of the National Institute of Genetics, 3:62–63, 1953.

[8] Motoo Kimura et al. Evolutionary rate at the molecular level. Nature, 217(5129):624–626, 1968.

[9] Motoo Kimura and George H Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics, 49(4):561, 1964.

[10] Takeo Maruyama. Genetic correlation in the stepping stone model with non-symmetrical migration rates. Journal of Applied Probability, 6(3):463–477, 1969.

[11] Takeo Maruyama. Analysis of population structure. Annals of human genetics, 34(2):201–219, 1970.

[12] Takeo Maruyama. On the rate of decrease of heterozygosity in circular stepping stone models of populations. Theoretical population biology, 1(1):101–119, 1970.

[13] Jennifer E Purcell. Jellyfish and ctenophore blooms coincide with human proliferations and environmental perturbations. Annual Review of Marine Science, 4:209–235, 2012.

[14] Timothy Sauer. Numerical Analysis. Pearson, Boston, MA, 2006.

[15] Rudolf S Scheltema. On dispersal and planktonic larvae of benthic invertebrates: an eclectic overview and summary of problems. Bulletin of Marine Science, 39(2):290–322, 1986.

[16] Wilhelm Weinberg. Über vererbungsgesetze beim menschen. Molecular and General Genetics MGG, 1(1):440–460, 1908.

[17] George H Weiss and Motoo Kimura. A mathematical analysis of the stepping stone model of genetic correlation. Journal of Applied Probability, pages 129–149, 1965.

[18] Sewall Wright. Isolation by distance. Genetics, 28(2):114, 1943.

[19] Ryo Yamaguchi and Yoh Iwasa. First passage time to allopatric speciation. Interface focus, 3(6), 2013.

Appendix A

General Procedure for Numerical Simulation

A.1 Gene Flow Process

The gene flow process observed in Kimura and Weiss's one-dimensional stepping stone model is readily simulated by tracking the movement and reproduction of individuals. The population of interest is subdivided onto a ring consisting of a large constant number of discrete islands, which are connected by one-step linear migration routes (Figure 3.1). These islands are initially populated with individuals represented by the digits 1 and 0 (to designate whether that individual has or does not have the allele in question, respectively).

After setting up the population structure and defining the migration rates, we begin to simulate the evolution of the population over the course of many generations. In the migration step, a loop over all individuals allows the opportunity for them to migrate one island to the left or right, to stay put on the current island, or to engage in long-range dispersal to any island chosen at random. The probability of each action is preset by the user. After migration, the number of individuals on a certain island may exceed or fall short of that island's designated population size/carrying capacity n_i. The reproduction step remedies this: a new generation is formed by choosing (with replacement) n_i "parent" individuals to pass on their allele value to their offspring. These migration and reproduction processes then repeat for as many generations as desired.

All code was written in C++.

A.2 Notable Algorithms

Migration Process

```
//Migration Loop:
//Inner loop moves the individuals on given island; outer loop cycles
//through all islands.
//For each individual,determine whether they stay put,move left or
```

```
//right,or move "long-distance" (to any island at random).
//Assign (push_back) each individual to their new island in the
//postMigration vector.

if(g % printWarning == 0){cout <<"Migrating... ";}

for(currIsle = 0; currIsle <numIslands; currIsle++)
{

for(indiv = 0; indiv <initialPerIsland; indiv++)
{
//A random "Choice Value" determines where each individual moves (or
stays):
double choice = randVal();

if(choice <= m_infinity){
//Long range migration
//Move to ANY island at random:

newIsle = rand() % numIslands;

}
else if(choice >m_infinity && choice <= m_infinity + (m/2.0)){
//Migrate left

if(USE_BINOMIAL == 0){
numStepsToMove = 1;
}
else{
//Function: int binomial(double p,int n)
numStepsToMove = binomial(binPVal,MAX_N -1 ) +1;
}

if(USE_CIRCULAR == 1){
//Circular model
newIsle = (currIsle <= (numStepsToMove -1))? (numIslands
-numStepsToMove + currIsle) :(currIsle-numStepsToMove);
}
else{
//Linear model with capped ends
newIsle = (currIsle <= (numStepsToMove -1))? currIsle :
(currIsle-numStepsToMove);
}
```

```
}
else if(choice >m_infinity + (m/2.0) && choice <= m_infinity + m){
//Migrate right

if(USE_BINOMIAL == 0){
numStepsToMove = 1;
}
else{
//Function: int binomial(double p,int n)
numStepsToMove = binomial(binPVal,MAX_N -1 ) +1;
}

if(USE_CIRCULAR == 1){
//Circular model
newIsle = (currIsle >= (numIslands-numStepsToMove))? (currIsle
-numIslands + numStepsToMove) : (currIsle+numStepsToMove);
}
else{
//Linear model with capped ends
newIsle = (currIsle >= (numIslands-numStepsToMove))? currIsle :
(currIsle+numStepsToMove);
}

}
else{
//Stay put on current island
newIsle = currIsle;

}//end if/else

migrationPopulation[newIsle].push_back(islandPopulation[currIsle][indiv]);


}//end for loop (individuals)

}//end for loop (islands)
```

Reproduction Process

```
//Reproduction Loop:
```

```
//For each island,randomly choose (with replacement) a "parent"
//individual from that island's post-migration state.
//Copy the parent's allele status (0 or 1) into the new postRepro
island. //This entry is the "child".
//Repeat until you have reached the desired number of individuals
//(population size) for the given island.

if(g % printWarning == 0){cout <<"Reproducing... ";}

for(currIsle = 0; currIsle <numIslands; currIsle++)
{
tally = 0;
for(i = 0; i <initialPerIsland; i++)
{
parent = rand() % migrationPopulation[currIsle].size();
islandPopulation[currIsle][i] = migrationPopulation[currIsle][parent];
tally += migrationPopulation[currIsle][parent];

}//end for loop (single island)


freqs[currIsle] = ((double) tally) / ((double) initialPerIsland);

migrationPopulation[currIsle].resize(0);

}//end for loop (all islands)
```