Estimating and modeling gene flow for a spatially distributed species

By

T. Burr[1]

and

T. V. Kurien[2]

Department of Statistics
Florida State University

January 1991
FSU Technical Report Number M 837
U.S. Army Technical Report Number D 116
AFOSR Technical Report Number 91-255

Key Words: Migration rates, eigenvalues, allele, Stepping-Stone model, Island model, Gaussian model.

[1] Research supported by Florida State University

[2] Research supported by USARO Grant No. DAAL03-90-G-0103 and by AFOSR Grant No. 91-0048.

*E*stimating and modeling gene flow for a spatially distributed species.

By
T. Burr and T. V. Kurien
Department Of Statistics
Florida State University

Abstract

This paper models the genetic behavior of a large population of individuals which is divided into colonies. We are studying the relative frequency of an allele $A1$ at a specific locus and over all the colonies. The effects of various migration patterns across colonies on these relative freqencies are studied. We set down a joint distribution for the relative frequencies of $A1$ at the colonies. This joint distribution is Gaussian and allows us to estimate parameters that describe the extent of genetic exchange across colonies. These parameters are called migration rates. The Gaussian models fits well to observed data and is very easy to simulate. The model can be extended without much difficulty to describe various mating patterns across colonies.

1 Introduction

Population geneticists have many models to study the effect of geographical subdivision on the evolution of a species.

Consider a large population of individuals of a particular species which is to some extent subdivided into colonies. Complete subdivision means that each colony is isolated. At the other extreme is no subdivision. This means that all adults in the entire population are equally likely to mate with all other adults of the opposite sex

in the population. It is believed that many species follow mating patterns somewhere between these two extremes. To set the stage for the mathematical models, the necessary genetical terms are collected here. Most organisms are *diploid*, having chromosomes in pairs, one inherited from each parent's *gamete* (sperm or egg) cell. The *genotype* of a diploid individual is the specification of all of its chromosome pairs. It is sometimes sufficient to model a diploid species as if it were haploid. *Haploid* individuals have only one of each type of chromosome. We think of a chromosome as a long string of symbols from the four-letter DNA alphabet representing the four different nucleotides of DNA. At a certain place on a chromosome (referred to as a *locus*) is a meaningful string of several hundred symbols called a *gene*. Typically there are many loci on a chromosome. The possible messages that could be written are called the *alleles* of that locus. Though the characteristics of an individual depend in a complicated way on all loci, it is informative to study a single locus. If this locus is a string of 200 letters there are 4^{200} alleles. At the present time it is common to detect alleles by *electrophoresis* which typically is able to detect only a few alleles from the essentially infinite number of possible alleles at the DNA level. If only two alleles can be distinguished at a locus, say A1 and A2, then the three possible genotypes for that particular locus are A1A2, A1A1, and A2A2, which are referred to as *heterozygote*, *homozygote*, and *homozygote*, respectively. *Natural selection* is in effect if the different genotypes have different probabilities of contributing to the gene pool in the next generation. If, for example, heterozygotes have a larger probability of surviving to adulthood and contributing to the gene pool of the next generation, then heterozygotes are said to have a greater fitness. An allele that is not affected by selection is called *neutral*. If an allele is altered by a copying error or environmental stimulus, the resulting allele is called a *mutant*. The present work considers loci with two neutral alleles. Note that if there are more than two alleles at the locus under study, then the alleles could be grouped into two groups.

Two simple models that represent extremes in migration habits are the Island Model and the Stepping-Stone Model. (Wright, 1931 and Kimura, 1953) In the Island Model all colonies are equally likely to exchange migrants with all other colonies. In the Stepping-Stone Model each colony exchanges migrants with only its nearest neighboring colonies. These two extremes will be considered here and in the context of the Gaussian Model will be referred to as the full-neighbor and 4-neighbor models.

Kimura (1953) proposed a migration pattern where migration is from nearby colonies. This is called the Stepping-Stone model. In the two-dimensional Stepping-Stone model each colony is located at a grid point in an n by n lattice. Identify the location of a colony by $\mathbf{i}$ where $\mathbf{i} \in \{(0,0),(0,1),\ldots(n-1,n-1)\}$. The version studied here is the following: Each colony maintains a constant population size N, with 2 alleles at each locus. Since there is a finite number of colonies, a joint non-trivial stationary distribution of the relative allele frequency of A1 in each colony is possible only if there is reversible mutation or migration from a constant outside source.

Assume here that there is an outside source, say a mainland population with constant A1 allele relative frequency p_M. Denote the relative frequency of allele A1 in the colony at $\mathbf{i}$ by $p_{\mathbf{i}}$. To specify a particular migration pattern, assume that the colony at $\mathbf{i}$ replaces a fraction, m, of its population with immigrants from the four neighbors. $\{(i_1-1,i_2),(i_1+1,i_2),(i_1,i_2-1),(i_1,i_2+1)\}$ and also replaces a small fraction, m_l, of its population with immigrants from a large mainland population. Here, addition is modulo n to avoid edge effects. For example, the four neighbors of $(n-1,n-1)$ are $\{(n-2,n-1),(0,n-1),(n-1,n-2),(n-1,0)\}$. Assume that all individuals produce many gametes (sperm or egg cells) and the stochastic component of how one generation leads to the next is due to the binomial sampling involved in reducing each population to size N.

Attempts to obtain the steady-state distribution of the relative frequency of al-

lele A1 in a single colony have not been successful. However, a Beta distribution appears to give a good approximation in simulation studies (Maruyama, 1977). Weiss and Kimura (1965) obtained an expression for the stationary correlation between p_i and p_j without attempting to approximate the joint stationary distribution.

A rather different approach is to approximate the joint steady state distribution of $\{p_i : i \in (i_1, i_2) : 0 \leq i_1 \leq n-1, 0 \leq i_2 \leq n-1\}$.

There has been no published attempt to approximate this joint distribution; however, in the full-neighbor case, as $n \to \infty$, one could use standard diffusion theory to arrive at an approximating stationary distribution. (See, for example, Mathematical Population Genetics.) The approximating stationary distribution would in that case be greatly simplified since one could take the p_i to be independent beta random variables, with parameters determined by the number of migrants exchanged per generation and the mean allele relative frequency.

To see that the Stepping-Stone model gives rise to a Markov Random Field, write p_i for the relative frequency of allele A1 in colony i. Then the conditional distribution of p_i given the relative frequency of A1 in all the colonies is the same as the conditional distribution given the relative frequency of A1 in only the neighbors of i. Since $0 \leq p_i \leq 1$, it will be necessary to transform the $\{p_i\}$ in order to use the Gaussian model. However, for some values of the parameters, the Gaussian model should fit the raw data. Under the Gaussian model, to be given in section 2, the conditional distribution of X_i given the relative frequencies of A1 in its neighbors will be Gaussian. Empirically, it appears that a transformation will be necessary except if μ is near .5 and the number of migrants exchanged is fairly large, say 2 or more. The choice of transformation was made by considering that the marginal distributions do appear to be approximately beta. Typical transforms from a beta distribution to approximate normality include the logit and probit transforms, which are $\log(p/(1-p))$ and $\Phi^{-1}(p)$, respectively, where Φ is the standard normal cumulative distrubtion function. For some values of the parameters, these transforms

might be improved by first raising p to some power. Therefore, the Gaussian Model has been fit to $x = \log(p/(1-p))$,
$x = \log(p^\lambda/(1-p^\lambda))$, $x = \Phi^{-1}(p)$, and to $x = \Phi^{-1}(p^\lambda)$. Good values for λ seem to be 1.5 to 2.

2 The Gaussian Model

Notation

Let:

$L_n^2 = \{(i_1, i_2) : 0 \le i_1 \le n-1, 0 \le i_2 \le n-1\}$ be an n by n array.

N_0 be a neighborhood of 0. For example, one could take N_0 to be $\{(0,1),(1,0),(0,-1),(-1,0)\}$. This is the usual 4-neighbor lattice.

$\mathbf{i} + \mathbf{j} = (i_1 \oplus j_1, i_2 \oplus j_2)$ where $\oplus$ is addition modulo n.

Points in L_n^2 are 2 component vectors *generically* denoted by $\mathbf{i}, \mathbf{j}, \mathbf{k}$, and $\mathbf{l}$. Let $\mathbf{1}$ denote the vector $(1,1) \in L_n^2$.

Assume that the joint stationary density of $X_\mathbf{i}$ is the multivariate normal (MVN) density:

$$f(\mathbf{x}) = |A|^{1/2} (2\pi)^{-n^2/2} \exp\{-\frac{(\mathbf{x} - \mu\mathbf{1})^T A (\mathbf{x} - \mu\mathbf{1})}{2}\}. \tag{2.1}$$

Note that the mean of $\mathbf{x}$ is $\mu\mathbf{1}$ and the covariance matrix of $\mathbf{x}$ is A^{-1}.

To capture the nearest-neighbor migration patterns, rewrite $f(\mathbf{x})$ as:

$$f(\mathbf{x}) = |A|^{1/2} (2\pi)^{-n^2/2} \exp\{-1/2 \sum_{\mathbf{i} \in L_n^2} \sum_{\mathbf{j} \in N_0} c(\theta_1/d, \mathbf{j})(x_\mathbf{i} - x_{\mathbf{i}+\mathbf{j}})^2 - \theta_2/2 \sum_{\mathbf{i} \in L_n^2} (x_\mathbf{i} - \mu)^2\}. \tag{2.2}$$

Here, $c(\theta_1/d, \mathbf{j})$ determines the amount of migration among neighbors. Large values of θ_1 correspond to large migration rates between neighbors. Large values of

θ_2 correspond to large migration rates from the constant outside source, or to large mutation rates. The index j allows for the migration rate to depend on direction, and d is the number of neighbors. Assuming d is known, one goal will be to estimate θ_1 and θ_2.

From (1.1) and (1.2) it can be shown that:

$$\begin{aligned} A_{0,0} &= A_{\mathrm{i,i}} = \theta_2 + 2 \sum_{\mathrm{j} \in N_0} c(\theta_1/d, \mathrm{j}) \\ A_{\mathrm{i,i+j}} &= -(c(\theta_1/d, \mathrm{j}) + c(\theta_1/d, \mathrm{j})) \text{ if } \mathrm{j} \in N_0 \\ A_{\mathrm{i,i+j}} &= 0 \text{ if } \mathrm{j} \notin N_0. \end{aligned} \tag{2.3}$$

The eigenvalues of A are:

$$\lambda_{\mathrm{k}}^n = \theta_2 + 2 \sum_{\mathrm{j} \in N_0} c(\theta_1/d, \mathrm{j})(1 - \cos(2\pi < \mathrm{j}, \mathrm{k} > /n)). \tag{2.4}$$

Although λ_{k}^n depends on n we shall write λ_{k}^n as λ_{k} from now on for notational convenience. Also, let Λ be the diagonal matrix with diagonal entries λ_{k}.

3 The Stepping-Stone model

For the Stepping-Stone model we shall assume isotropic migration, by putting $c(\theta_1/4, \mathbf{j}) = c(\theta_1/4, -\mathbf{j}) = \theta_1/4$. The Stepping-Stone model accounts for local migration by taking $N_0 = \{(0,1),(1,0),(0,-1),(-1,0)\}$.

Then (2.3) and (2.4) simplify to:

$$\begin{aligned} A_{0,0} &= A_{\mathbf{i},\mathbf{i}} = \theta_2 + 2\theta_1 \\ A_{\mathbf{i},\mathbf{i}+\mathbf{j}} &= -\theta_1/2 \ \forall \mathbf{i} \in L_n^2 \ \text{if } \mathbf{j} \in N_0 \\ A_{\mathbf{i},\mathbf{i}+\mathbf{j}} &= 0 \text{ if } \mathbf{j} \notin N_0 \end{aligned} \tag{3.1}$$

$$\lambda_{\mathbf{k}} = \theta_2 + \theta_1((1 - \cos(2\pi k_1/n) + (1 - \cos(2\pi k_2/n)). \tag{3.2}$$

The model accounts for long distance migration (Island model) if we take $N_0 = L_n^2 - (0,0)$, which is the full-neighbor version. Then (2.3) and (2.4) simplify to:

$$\begin{aligned} A_{0,0} &= A_{\mathbf{i},\mathbf{i}} = \theta_2 + 8\theta_1/(n^2-1) \\ A_{\mathbf{i},\mathbf{i}+\mathbf{j}} &= -2\theta_1/(n^2-1) \ \forall \mathbf{i} \in L_n^2 \ \text{if } \mathbf{j} \in L_n^2 - (0,0) \end{aligned} \tag{3.3}$$

$$\begin{aligned} \lambda_{\mathbf{k}} &= \theta_2 + 2n^2\theta_1/(n^2-1) \text{ if } \mathbf{k} \in L_n^2 - (0,0) \\ \lambda_{0,0} &= \theta_2. \end{aligned} \tag{3.4}$$

In order to generate data from the density (2.2), first generate n^2 independent standard normal random variables, say $\mathbf{Z} \sim \mathrm{N}(0,I)$, where I is the n^2 by n^2 identity matrix. To obtain $\mathbf{X} \sim N(0, A^{-1})$, use the fact that if $\mathbf{Y} \sim N(0,\Sigma)$ then $C\mathbf{Y} \sim N(0, C\Sigma C^T)$ for C an appropriately dimensioned matrix of constants. Now use the

spectral decomposition of the covariance matrix A^{-1} in the usual way to find C such that $CC^T = A^{-1}$. This gives the following prescription for generating the data. The observation at the location $\mathbf{l} = (l_1, l_2)$ is generated by:

$$X_{\mathbf{l}} = \mu + n^{-2} \sum_{\mathbf{k} \in L_n^2} Z_{\mathbf{k}} \sum_{\mathbf{j} \in L_n^2} \frac{\cos(2\pi(<\mathbf{l},\mathbf{j}> + <\mathbf{k},\mathbf{j}>)/n)}{\sqrt{\lambda_{\mathbf{j}}}}. \tag{3.5}$$

This means that an observation from the steady-state distribution of an n by n array requires simply the generation of n^2 standard normal random variables, and performing the indicated summation. This is much faster than simulating the migration and reproduction (random sampling of gametes) pattern for many (approximately 100) generations until stationarity is reached.

One of the main results of previous work with the stepping stone model is how the covariance between p_0 and p_l depends on the dimension of the habitat (linear, in the plane, or in three dimensions), the migration pattern, and the migration rate. Maruyama, Kimura and Weiss all used recursion equations to model the way that the relative frequencies change each generation and solved for the stationary covariance, $\mathrm{cov}(p_0, p_l)$. An attractive feature of the Gaussian model is that $\mathrm{cov}(X_0, X_l)$ is easier to compute than it is when working with the recursion equations. From (3.5) it can be shown that:

$$\mathrm{cov}(X_0, X_{\mathbf{l}}) = n^{-4} \sum_{\mathbf{k} \in L_n^2} \sum_{\mathbf{j},\mathbf{s} \in L_n^2} \frac{\cos(2\pi(<\mathbf{j},\mathbf{k}> + <\mathbf{s},\mathbf{k}+\mathbf{l}>)/n)}{\sqrt{\lambda_{\mathbf{j}}\lambda_{\mathbf{s}}}}. \tag{3.6}$$

It is known that migration rates among partially isolated colonies need not be very large to prevent genetic diversity. However, if migration tends to be from nearest-neighboring colonies rather than the entire population, then there is greater potential for genetic diversity (Crow and Aoki, 1982). The Gaussian model has this same feature. One way to see this is to solve for the variance of the equilibrium distribution of the relative frequency of allele A1 in any of the identical colonies. Let $\lambda_{\mathbf{k}}$ be given by equation 2.4. Then it follows from the spectral representation

of the variance-covariance matrix that the variance of the relative frequency of A1 in each colony is:

$$\text{var}(\bar{X}) = n^{-2} \sum_{k \in L_n^2} 1/\lambda_k. \tag{3.7}$$

This provides an easy proof that var(X) is less in the full-neighbor model than in the 4-neighbor model, which is intuitively expected. To see this, first note that $\lambda_{0,0} = \theta_2$ in both the full-neighbor and the 4-neighbor models. Next, it can be shown that all remaining eigenvalues are the same in the full-neighbor case, but are not all the same in the 4-neighbor case. Also, the sum of the eigenvalues can be shown to be the same in both cases. Then, since the harmonic mean of a collection of unequal numbers is necessarily less than the arithmetic mean of that same collection of numbers, the result follows.

4 Estimation

The three parameters can be estimated by maximum likelihood using the density in (2.1). The following is specifically for the 4-neighbor model but could easily be modified for any neighborhood structure.

The MLE $\hat{\mu}$ of μ is the sample mean $\bar{x}$.

Let $\omega_k = 1 - \cos(2\pi k_1/n) + 1 - \cos(2\pi k_2/n)$, and let $d = 4$. Note that $\lambda_k = \theta_2 + \theta_1 \omega_k$ (see 3.2).

The likelihood equations for θ_1 and θ_2 are:

$$n^{-2} \sum_{k \in L_n^2} \frac{4\omega_k}{\widehat{\theta_2} + \hat{\theta}_1 \omega_k} = n^{-2} \sum_{i \in L_n^2} \sum_{j \in N_0} (x_i - x_{i+j})^2 \tag{4.1}$$

$$n^{-2} \sum_{k \in L_n^2} \frac{1}{\widehat{\theta_2} + \hat{\theta}_1 \omega_k} = n^{-2} \sum_{i \in L_n^2} (x_i - \bar{x})^2 \tag{4.2}$$

Theorem 1.

The RHS of (1.1) is equal in distribution to $n^{-2}\sum_{k\in L_n^2} 4\omega_k\lambda_k^{-1}\chi_1^2$. The RHS of (1.2) is equal in distribution to $n^{-2}\sum_{k\neq(0,0)}\lambda_k^{-1}\chi_1^2$.

Proof

The RHS of (1.1) is a quadratic form, so it can be expressed as X^TBX, where

$$\begin{aligned} B_{0,0} &= B_{i,i} &&= 4 \\ B_{0j} &= B_{i,i+j} &&= 0 \text{ for } j \notin N_0 \\ & && = 1 \text{ for } j \in N_0. \end{aligned}$$

Let the covariance matrix of X be Σ, and let the matrix with the eigenvectors of Σ be denoted P^T. Let $Y = P^TB^{1/2}X$. Then $Y^TY = X^TBX$, Y is a vector of independent, mean 0 normal random variables, and Y_k has variance $4\omega_k/\lambda_k$. The result follows by observing that a χ^2 random variable can be generated by squaring a standard normal random variable.

To find the distribution of the RHS of (1.2), let $Y = P^TX$ so that $Y^TY = X^TX$, and $Y \sim N(P^T\mu, \Lambda^{-1})$. Therefore, $\sum_{k\neq0,0} x_i^2 \stackrel{d}{=} \theta_2^{-1}\chi^2(\theta_2 n^2\mu) + \sum_{k\neq0,0}\lambda_k^{-1}\chi^2$. Then, $n^{-2}\sum_{i\in L_n^2}(x_i-\overline{x})^2 = n^{-2}\sum_{i\in L_n^2} x_i^2 - \overline{x}^2 \stackrel{d}{=} n^{-2}\sum_{i\in L_n^2} y_i^2 - \overline{x}^2$. The result follows by observing that $y_{0,0} = n\overline{x}$. □

Now let a_1 and a_2 be defined by:

$$\begin{aligned} a_1 &= \lim_{n\to\infty} n^{-2}\sum_{k\in L_n^2} 4\omega_k/\lambda_k \\ &= \int_0^1\int_0^1 4(2-\cos(2\pi x)-\cos(2\pi y))/(\theta_2+\theta_1(2-\cos(2\pi x)-\cos(2\pi y)))\,dxdy \;<\; \infty \\ a_2 &= \lim_{n\to\infty} n^{-2}\sum_{k\in L_n^2} 1/\lambda_k \\ &= \int_0^1\int_0^1(\theta_2+\theta_1(2-\cos(2\pi x)-\cos(2\pi y)))^{-1}\,dxdy \;<\infty. \end{aligned}$$

By a version of the CLT, the RHS of 1.1 is $AN(a_1, \sigma_n^2)$, where $\sigma_n^2 \to 0$. Similarly, the RHS of 1.2 is AN with mean a_2 and a variance that goes to zero as $n \to \infty$.

The asymptotic distribution of the MLE's of θ_1 and θ_2 has been established by first noting that the likelihood can be written in terms of independent random variables by using $Y = P^TX$, where P^T is the matrix of eigenvectors of the covariance

matrix of X. As in the proof of theorem 1, this follows because $Y \sim N(P^T\mu, \Lambda^{-1})$, making Y a vector of n^2 independent normal random variables with different variances, one with mean $n\mu$, and all others with mean 0. The likelihood is then $\prod_{k\in L_n^2} f_k(x_k,\theta)$, where $f_k(x_k,\theta) = (\lambda_k/(2\pi))^{1/2}\exp(-\lambda_k x_k^2/2)$ for all k except $k = (0,0)$, since $Y_{0,0}$ has nonzero mean, with density $(\theta_2/(2\pi))^{1/2}\exp(-\theta_2(x_{0,0}-n\mu)^2/2)$.

The likelihood equations for Y are the same as those for X. The asymptotic normality of $\widehat{\theta} = (\widehat{\theta_1}, \widehat{\theta_2})$ now follows from theorem 2(iv) of Bradley and Gart (1962).

Specifically, let c_1, c_2, and c_3 be defined by:

$$\begin{aligned} 2c_1 &= \lim_{n\to\infty} n^{-2} \textstyle\sum_{k\in L_n^2} \omega_k^2/\lambda_k^2 \\ &= \int_0^1\int_0^1 (2-\cos(2\pi x)-\cos(2\pi y))^2/(\theta_2+\theta_1(2-\cos(2\pi x)-\cos(2\pi y)))^2\, dxdy \\ 2c_2 &= \lim_{n\to\infty} n^{-2} \textstyle\sum_{k\in L_n^2} \omega_k/\lambda_k^2 \\ &= \int_0^1\int_0^1 (2-\cos(2\pi x)-\cos(2\pi y))/(\theta_2+\theta_1(2-\cos(2\pi x)-\cos(2\pi y)))^2\, dx\, dy \\ 2c_3 &= \lim_{n\to\infty} n^{-2} \textstyle\sum_{k\in L_n^2} \lambda_k^{-2} \\ &= \int_0^1\int_0^1 (\theta_2+\theta_1(2-\cos(2\pi x)-\cos(2\pi y))^{-2} dxdy\ . \end{aligned}$$

It follows that c_1, c_2, and c_3 are finite.

Theorem 2.

$n(\widehat{\theta}-\theta) \xrightarrow{d} N(0, I^{-1})$, where $I = \begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix}$.

In particular, $n(\widehat{\theta_1}-\theta_1) \xrightarrow{d} N(0, c_3(c_1c_3-c_2^2)^{-1})$ and $n(\widehat{\theta_2}-\theta_2) \xrightarrow{d} N(0, c_1(c_1c_3-c_2^2)^{-1})$.

Theorem 2 (iv) of Bradley and Gart is appropriate when the number of populations sampled increases as the sample size increases. The proof will consist of

verifying the conditions of their theorem. In order to conform more with their notation, the boldface vector notation will not be used while verifying the conditions. Since the data is a single observation from each site in an n by n array, the sample size is n^2 instead of the usual n. Assume that one sample is taken from each density $f_i(x_i, \theta)$. The joint likelihood is then $\prod_{i=1}^{n^2} f_i(x_i, \theta)$. In general, it is not necessary that each f_i depend on both θ_1 and θ_2, but in the present case each f_i does depend on both θ_1 and θ_2.

Let $\theta = (\theta_1, \theta_2)$, $\Omega = (0, \infty) \times (0, \infty)$. The conditions to verify are:

1(a). For almost all $x_i \in \Re$ and for all $\theta \in \Omega$, $\partial \log f_i/\partial\theta_1$, $\partial \log f_i/\partial\theta_2$, $\partial^2 \log f_i/\partial\theta_1^2$, $\partial^2 \log f_i/\partial\theta_2^2$, $\partial^2 \log f_i/\partial\theta_1\partial\theta_2$, $\partial^3 \log f_i/\partial\theta_1^2\partial\theta_2$, $\partial^3 \log f_i/\partial\theta_1\partial\theta_2^2$ exist for $i = 1, \ldots, n^2$.

1(b). For all f_i, for almost all x_i, and for every $\theta \in \Omega$,
$|\partial f_i/\partial\theta_1| < F_{i1}(x_i)$, $|\partial f_i/\partial\theta_2| < F_{i2}(x_i)$, $|\partial^2 f_i/\partial\theta_1^2| < F_{i11}(x_i)$, $|\partial^2 f_i/\partial\theta_2^2| < F_{i22}(x_i)$,
$|\partial^2 f_i/\partial\theta_1\partial\theta_2| < F_{i12}(x_i)$, $|\partial^3 \log f_i/\partial\theta_1^2\partial\theta_2| < H_{i112}(x_i)$, $|\partial^3 \log f_i/\partial\theta_1\partial\theta_2^2| < H_{i122}(x_i)$,
where $F_{ir}(x_i)$ and $F_{irs}(x_i)$ are integrable over $\Re$ and $\int_{\Re} H_{irst}(x_i) f_i \, dx_i < M_i$,
with the M_i finite positive constants, for $i = 1, \ldots n^2 - 1$, and $r, s, t = (1, 2)$.

2(a). For the sequence of density functions $\{f_i\}$, $\sum_{i=1}^{n^2} \int_{D_{1i}} f_i \, dx_i = o(1)$,
where $D_{1i} = \{|\partial \log f_i/\partial\theta_r| > n^2\}$ and $\sum_{i=1}^{n^2} \int_{D_{2i}} (\partial \log f_i/\partial\theta_r)^2 f_i \, dx_i = o(n^4)$,
where $D_{2i} = \{|\partial \log f_i/\partial\theta_r| < n^2\}$ for $r = 1, 2$.

2(b). For the sequence of density functions $\{f_i\}$, $\sum_{i=1}^{n^2} \int_{D_{3i}} f_i \, dx_i = o(1)$,
where $D_{3i} = \{|\partial^2 \log f_i/\partial\theta_r\partial\theta_s| > n^2\}$, $\sum_{i=1}^{n^2} \int_{D_{4i}} |\partial^2 \log f_i/\partial\theta_r\partial\theta_s| f_i \, dx_i = o(n^4)$,
where $D_{4i} = \{|\partial^2 \log f_i/\partial\theta_r\partial\theta_s| < n^2\}$, and
$\lim_{n^2 \to \infty} n^{-2} \sum_{i=1}^{n^2} \int_{\Re} |\partial^2 \log f_i/\partial\theta_r\partial\theta_s| f_i \, dx_i = J_{rs}(\theta)$ exists, and that $J_{rs}(\theta)$ be positive definite with finite determinant for $r, s = 1, 2$.

2(c). Referring to H_{irst} and M_i in 1(b). $\sum_{i=1}^{n^2} \int_{D_{5i}} f_i \, dx_i = o(1)$, where $D_{5i} = \{H_{irst} > n^2\}$ and $n^{-1} \sum_{i=1}^{n^2-1} M_i < M$ with M a finite positive constant.

3(a). For every $\epsilon > 0$, $\lim_{n-\infty} n^{-2} \sum_{i=1}^{n^2} \int_{D_{7i}} \sum_{r=1}^{2} (\partial \log f_i / \partial \theta_r)^2 f_i \, dx_i = 0$, where $D_{7i} = \{\sum_{r=1}^{2} (\partial \log f_i / \partial \theta_r)^2 > \epsilon n^2.\}$

Verification of the conditions will now be given.
Recall that for $i \neq (0,0)$, $f_i(x_i, \theta) = (\lambda_i / 2\pi)^{1/2} \exp(-\lambda_i x_i^2) \omega_i$,
and $\omega_i = 2 - \cos(2\pi i_1 / n) - \cos(2\pi i_2 / n)$. The $i = (0,0)$ observation has non-zero mean and should be treated separately since the derivatives needed will be slightly different. It is easily verified that the (0,0) observation poses no difficulty, so for brevity, will not be treated here.

1(a). Clearly these derivatives exist, and are recorded here for later use.
$\partial \log f_i / \partial \theta_1 = (\omega_i / 2)(1/\lambda_i - x_i^2)$, $\partial \log f_i / \partial \theta_2 = (1/2)(1/\lambda_i - x_i^2)$,
$\partial^2 \log f_i / \partial \theta_1^2 = -\omega_i^2 / (2\lambda_i^2)$, $\partial^2 \log f_i / \partial \theta_2^2 = -1/(2\lambda_i^2)$, $\partial^2 \log f_i / \partial \theta_1 \partial \theta_2 = -\omega_i / (2\lambda_i^2)$,
$\partial^3 \log f_i / \partial \theta_1^2 \partial \theta_2 = \omega_i^2 / \lambda_i^3$, $\partial^3 \log f_i / \partial \theta_1 \partial \theta_2^2 = \omega_i / \lambda_i^3$

1(b). $\partial f_i / \partial \theta_2 = (\lambda_i^{-1/2} - \lambda_i^{1/2}) \exp(-\lambda_i x_i^2 / 2) / (2(2\pi)^{1/2})$
Since $0 \leq \omega_i \leq 4$, and $\theta_2 \leq \lambda_i \leq \theta_2 + 4\theta_1$,
$|\partial \log f_i / \partial \theta_2| < (\theta_2^{-1/2} + (\theta_2 + 4\theta_1)^{1/2}) \exp(-\lambda_i x_i^2 / 2) / (2(2\pi)^{1/2}) = F_{i2}(x_i)$, which is integrable over $\Re$.

Since $\partial f_i / \partial \theta_1 = \omega_i \partial f_i / \partial \theta_2$, let $F_{i1}(x_i) = 4F_{i2}(x_i)$, which is integrable over $\Re$.

Next,

$$\begin{aligned} \partial^2 f_i / \partial \theta_2^2 &= -(\lambda_i^{-3/2} + \lambda_i^{-1/2} + x_i^2)(\lambda_i^{-1/2} - \lambda_i^{1/2})) \exp(-\lambda_i x_i^2 / 2) / (2(2\pi)^{1/2}) \\ |\partial^2 f_i / \partial \theta_2^2| &< (\theta_2^{-3/2} + \theta_2^{-1/2} + x_i^2 (\theta_2^{-1/2} + (\theta_2 + 4\theta_1)^{1/2})) \exp(-\lambda_i x_i^2 / 2) / (2(2\pi)^{1/2}) \\ &= F_{i11} \end{aligned}$$

Since $\partial^2 f_i/\partial\theta_1^2 = \omega_i^2 \partial^2 f_i/\partial\theta_2^2$, let $F_{i11} = 4F_{i22}$ and clearly both F_{i11} and F_{i22} are integrable over $\Re$.

Similarly,

$$\begin{aligned} \partial^2 f_i/\partial\theta_1\partial\theta_2 &= -\omega_i(2^{-1}\lambda_i^{-3/2} + 2^{-1}\lambda_i^{1/2} + (\lambda_i^{-1/2} - \lambda_i^{1/2})x_i^2/2)\exp(-\lambda_i x_i^2/2)/(2(2\pi)^{1/2}) \\ |\partial^2 f_i/\partial\theta_1\partial\theta_2| &< (\theta_2^{-3/2} + \theta_2^{-1/2} + (\theta_2^{-1/2} + (\theta_2 + 4\theta_1)^{1/2})x_i^2/2)\exp(-\lambda_i x_i^2/2)/(2\pi)^{1/2} \\ &= F_{i12}. \end{aligned}$$

And F_{i12} is integrable over $\Re$.

Next,

$$\partial^3 \log f_i/\partial\theta_1^2\partial\theta_2 = |\partial^3 \log f_i/\partial\theta_1^2\partial\theta_2| = \omega_i^2/\lambda_i^3 < 25/\theta_2^3 = H_{i112}$$
$$\partial^3 \log f_i/\partial\theta_1\partial\theta_2^2 = |\partial^3 \log f_i/\partial\theta_1^2\partial\theta_2| = \omega_i/\lambda_i^3 < 5/\theta_2^3 = H_{i122}$$

The condition $\int_{\Re} H_{irst} f_i dx_i < M_i$ is satisfied for all i for $r, s, t = 1, 2$ with $M_i = 25/\theta_2^2$.

2(a). $\partial \log f_i/\partial\theta_2 = (1/\lambda_i - x_i^2)/2$, so $\lim_{n\to\infty} P_i(D_{2i}) = 0$, where $P_i(D_{2i})$ is the measure associated with the random variable x_i. To see this, note that $\lambda_i^{1/2} x_i \sim N(0,1)$, so $x_i^2 \sim \lambda_i \chi_1^2$. Therefore, $P_i(|x_i^2 - 1/\lambda_i| > 2n^2) = P(|\chi_1^2 - 1| > 2(n^2)\lambda_i) \leq 1/(2\lambda_i^2 n^4) \leq 1/(2\theta_2^2 n^4)$, for all i, by Chebyshev's inequality.

This means that $\sum_{i=1}^{n^2} \int_{D_{2i}} f_i dx_i \leq n^2(\theta_2 + 4\theta_1)^{1/2}/((2\pi)^{1/2} 2\theta_2^2 n^4)$, which is o(1).

For the second part of 2(a), note that $(\partial f_i/\partial\theta_2)^2 = (x_i^4 - 2x_i^2/\lambda_i + 1/\lambda_i^2)/4$, so $\int_{D_{2i}} (\partial f_i/\partial\theta_2)^2 f_i dx_i < (3/\theta_2^2 + 2 + 1/\theta_2^2)/4 = b$, for all i, so the condition holds, since $n^2 b$ is $o(n^4)$.

Since $\partial \log f_i/\partial\theta_1 = \omega_i \partial \log f_i/\partial\theta_2$, these same conditions can be verified for $\partial \log f_i/\partial\theta_1$ in the same way.

2(b). For $r=1$ and $s=2$, $D_{3i} = \{|\partial^2 \log f_i/\partial\theta_1\partial\theta_2| > n^2\}$, and $|\partial^2 \log f_i/\partial\theta_1\partial\theta_2| = \omega_i/(2\lambda_i^2) \leq 2/\theta_2^2$, so $\lim_{n\to\infty} P_i(D_{3i}) = 0$. Therefore, for $n > 2/\theta_2^2$, all terms in the sum are zero, so the sum is zero.

By inspection of $\partial^2 \log f_i / \partial \theta_1^2$ and $\partial^2 \log f_i / \partial \theta_2^2$, similar results follow for each of these.

Since $\partial^2 \log f_i / \partial \theta_1^2$, $\partial^2 \log f_i / \partial \theta_2^2$, and $\partial^2 \log f_i / \partial \theta_1 \partial \theta_2$ are all bounded, the condition, $\sum_{i=1}^{n^2} \int_{D_{4i}} (\partial^2 \log f_i / \partial \theta_r \partial \theta_s)^2 \, f_i dx_i = o(n^4)$, is easily verified.

Also, $\partial^2 \log f_i / \partial \theta_1^2$, $\partial^2 \log f_i / \partial \theta_2^2$, and $\partial^2 \log f_i / \partial \theta_1 \partial \theta_2$ are all constants and $\lim_{n \to \infty} P_i(D_{4i}) = 1$, so the limit:

$\lim_{n \to \infty} \sum_{i=1}^{n^2} \int_{D_{4i}} - |\partial^2 \log f_i / \partial \theta_r \partial \theta_s| \, f_i dx_i = I_{rs}$ exists for $r, s = 1, 2$. The three elements of I were identified earlier as c_1, c_2, and c_3, with $2c_1 = \lim_{n \to \infty} n^{-2} \sum_{i=1}^{n^2} w_i^2 / \lambda_i^2$, $2c_2 = \lim_{n \to \infty} n^{-2} \sum_{i=1}^{n^2} \omega_i / \lambda_i^2$, $2c_3 = \lim_{n \to \infty} n^{-2} \sum_{i=1}^{n^2} 1 / \lambda_i^2$.

The matrix I must be positive definite, so it is necessary that $c_1 c_3 > c_2^2$. This follows from the Caucy-Schwarz inequality. Also, the determinant of I is finite, since c_1, c_2, and c_3 are all finite.

2(c). It has been shown that it suffices to take $H_{irst} = 25/\theta_2^3$ for $r, s, t = 1, 2$, so for $n > 25/\theta_2^3$, $P(D_{5i}) = 0$. Therefore, the conditions involving integrals over D_{5i} and D_{6i} hold. Also, it follows that $H_{irst} < M_i = 25/\theta_2^3$, so that $n^{-2} \sum_{i=1}^{n^2} M_i$ is bounded by $25/\theta_2^3$, which is finite.

3(a). This condition may be verified using the same approach as in 2(a), so will not be repeated here. □

5 Goodness of Fit

A Gaussian model has been proposed to explain the stationary distribution of the relative frequency of allele A1 in the Stepping-Stone model. The simplifications achieved justify its use, provided that it adequately describes the data and makes predictions that can be tested.

Two predictions of the Gaussian model that can be tested on real data are

the behavior of $\mathrm{cov}(X_0, X_1)$ and the behavior of $\mathrm{var}(\bar{X})$, where $\bar{X}$ is the sample average. Equation 3.6 gives the covariance between X_0 and X_1 and it can be shown that $\mathrm{var}(\bar{X}) = 1/(n^2\theta_2)$. These two predictions have been compared to several sets of data simulated from the Stepping-Stone model described in section 1. For the simulated data, it does appear that $\mathrm{var}(\bar{X}) \approx 1/(n^2\theta_2)$, In the Gaussian model, equation 3.6 implies that $\mathrm{cov}(\mathrm{X_0}, \mathrm{X_1})$ decreases with separation between 0 and 1 at a faster rate for small θ_1. The same is true for simulated data from the Stepping-Stone model.

Also, any test of multivariate normality can be applied to the simulated data. The Handbook of Statistics, Volume 1 has a few tests and reference to others. These tests have been applied on the marginal distributions from 200 observations of a 4 by 4 array at steady state: probit plots, D'Agostino's D (1971), Shapiro and Wilk's W (1965), skewness, kurtosis, and the Box and Cox transform (Gnanadesikan, 1977). Bivariate normality was checked using the Box and Cox transform. Multivariate normality (the joint distributions of all 16 random variables) was checked using multivariate tests of skewness, kurtosis (Mardia, 1970), Malkovich and Afifi's (1973) generalization of W, a χ^2_{16} probability plot of the Mahalanobis distances, and the associated Kolmogorov-Smirnoff (KS) test. Although the probability plot looked nearly linear, the uniform random variable associated with the KS test was .947. (Large values indicate lack of fit.) The Box and Cox transform was used on the raw data, say p_i, for i= 1,2,...,16, and on $p_i/(1-p_i)$. With the raw data, it was best not to transform and with $p_i/(1-p_i)$, the log transform was best. Regarding the marginals, 3 of the 16 skewness tests rejected and 5 of the 16 kurtosis tests rejected. Because p_M was taken to be .5, skewness is not expected but nonnormal kurtosis is a possibility. However, with 200 observations it would be surprising if none of the tests rejected the Gaussian assumption since it is an approximation to the unknown distribution. When p_M was taken to be .7, $\log(p_i/(1-p_i))$ has a skewness of about .4 and a kurtosis of about 3.4. This can be improved by using $\log(p_i^\lambda/(1-p_i^\lambda)$

with $\lambda = 1.5$ or 2. A better transform appears to be $\Phi^{-1}(p_i)$ or $\Phi^{-1}(p_i^{\lambda})$ where $\lambda = 1.5$ or 2. With $p_M = .7$ and using $\Phi^{-1}(p_i^{1.5})$, 200 observations from a 5 by 5 array appeared to be approximately MVN. Multivariate skewness and kurtosis were both within acceptable range, the χ^2_{25} plot of Mahalanobis distances looked very good, and the uniform random variable associated with the Kolmogorov-Smirnoff test was .0717. Only one of the 25 marginal tests of skewness and one of the 25 marginal tests of kurtosis were rejected at the .05 level. All of the above tests were with $m_l + m \geq 1$ to make the number of occurrences of $p_i = 1$ small. An observation of $p_i = 1$ was changed to $(N - .5)/N$ where N is the number per colony. Hereafter, it is assumed that $m_l + m \geq 1$.

Assuming that the Gaussian Model adequately describes the data, it is of interest to estimate θ_1 and θ_2 from simulated data and make predictions about the effect of changing from the 4-neighbor to the full-neighbor models. Using the $\Phi^{-1}(p_i^{1.5})$ transform, with $n = 10$, $m_l = .1$, and $m = 1.0$, 1.5, 2.0, 2.5, 3.0, θ_1 and θ_2 were estimated by maximum likelihood using a grid search to find good starting values, followed by the Newton-Raphson technique. The estimated values were used to predict the variance of the marginals for the full-neighbor model. Using $\theta_1/99$ versus $\theta_1/4$, for $m = 1.0$, 1.5, 2.0, 2.5, 3.0, the predicted values for the variance were .99, .73, .57, .47, and .41, respectively. As expected, these predictions are lower than those observed in the 4-neighbor case, which were 1.39, 1.06, .85, .71, and .63, respectively. A second simulation with the same values of m but using the full-neighbor migration pattern produced variances of 1.0, .73, .57, .45, and .39 for the 5 values of m. These are in good agreement with what had been predicted.

6 Summary

A Gaussian model has been fit to data simulated from the Stepping-Stone Model used in population genetics. Standard tests of multivariate normality on the transformed data suggest that the fit is acceptable, with the transform $\Phi^{-1}(p)$ or $\Phi^{-1}(p^{1.5})$ performing well. In addition, implications of the model do appear to hold for the transformed data.

The parameters of the Gaussian have been estimated by maximum likelihood, and the asymptotic distribution of the maximum likelihood estimators has been established.

The previous result by Kimura and Weiss for $\mathrm{cov}(X_0, X_1)$ has been derived (for the transformed data) without recourse to recurrence equations. A comparison of the Stepping-Stone neighborhood structure with the other extreme neighborhood structure, the Island model, has been made in terms of $\mathrm{var}(X_i)$. A more complete comparison using the total variation distance between the two joint distributions is possible.

Finally, the Gaussian model promises to have other applications. For example, the residuals from a regression model with spatial autocorrelation could be assumed to follow the Gaussian model. The parameter θ_1 provides a natural alternative in the hypothesis testing that is sometimes used in that context.

Bibliography

[1] D'Agostino. R.B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika* 58, 341-348.

[2] Bradley, R.A. and Gart, J. (1962). The Asymptotic properties of ML estimators when sampling from associated populations. *Biometrika* 49, 205-214.

[3] Ewens, W.J. (1979). Mathematical Population Genetics. Springer-Verlag.

[4] Feller, W. (1951). Diffusion processes in genetics. Proc. Second Berkeley Symp. Math. Stat. Prob. pp. 227-246.

[5] Gnanadesikan, R. (1977). Methods for statistical data analysis of multivariate observations. John Wiley & Sons.

[6] Grenander, U. and Sethuraman, J. (1985). Limit theorems in metric pattern theory. Unpublished.

[7] Hartl, D. and Clark, A. (1989) Theoretical Population Genetics. Sinauer Associates.

[8] Karlin, S. and Taylor, H. (1981). A Second Course in Stochastic Processes. Academic Press.

[9] Kimuara, M. (1953). "Stepping Stone" model of population. *Ann. Rept. Nat. Inst. Genet. Japan* 3, 62-63.

[10] Malkovich, J.F. and Afifi, A.A. (1973). On tests for multivariate normality. *JASA* 68, 176-179.

[11] Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 519-530.

[12] Maruyama, T. (1972). Distribution of gene frequencies in a geographically structured finite population. 1. Distribution of neutral genes and of genes with small effect. *Ann. Hum. Genet., London.* 35, 411-423.

[13] Maruyama, T. (1977). Lecture notes in biomathematics. Stochastic problems in population genetics. Springer-Verlag.

[14] Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality. *Biometrika* 52, 591-611.

[15] Weiss, G. and Kimura, M. (1965). A mathematical analysis of the stepping stone model of genetic correlation. *JAP* 2, 129-149.

[16] Wright, S. (1931). Evolution in Mendelian Populations. *Genetics* 16, 97-159.

[17] Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* 15, 323-354.

[18] Wright, S. (1969). Evolution and the genetics of populations, Vol. 2. The theory of gene frequencies. The University of Chicago Press 1969b.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): FSU Tech Report No. M837	5. MONITORING ORGANIZATION REPORT NUMBER(S): ARO 27868.6-MA
6a. NAME OF PERFORMING ORGANIZATION: Florida State University; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: U. S. Army Research Office
6c. ADDRESS (City, State, and ZIP Code): Department of Statistics, Florida State University, Tallahassee, Florida 32306	7b. ADDRESS (City, State, and ZIP Code): P. O. Box 12211, Research Triangle Park, NC 27709-2211
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: U. S. Army Research Office; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: DAAL03-90-G-0103
8c. ADDRESS (City, State, and ZIP Code): P. O. Box 12211, Research Triangle Park, NC 27709-2211	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. / PROJECT NO. DAAGL03 / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification): Estimating and modeling gene flow for a spatially distributed species

12. PERSONAL AUTHOR(S): T. Burr and T. V. Kurien

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	January 1991	21

16. SUPPLEMENTARY NOTATION: The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Migration rates, eigenvalues, allele, Stepping-Stone model, Island model, Gaussian model.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

SEE BACK.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: ☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION: Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) / 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

Abstract

This paper models the genetic behavior of a large population of individuals which is divided into colonies. We are studying the relative frequency of an allele $A1$ at a specific locus and over all the colonies. The effects of various migration patterns across colonies on these relative freqencies are studied. We set down a joint distribution for the relative frequencies of $A1$ at the colonies. This joint distribution is Gaussian and allows us to estimate parameters that describe the extent of genetic exchange across colonies. These parameters are called migration rates. The Gaussian models fits well to observed data and is very easy to simulate. The model can be extended without much difficulty to describe various mating patterns across colonies.