Analysis of Genetic Structure and Dispersal Patterns in a Population of Sea Beet

Jarle Tufto,* Alan F. Raybould,[†] Kjetil Hindar* and Steinar Engen[‡]

**Norwegian Institute for Nature Research, 7005 Trondheim, Norway, [†]Institute of Terrestrial Ecology, Furzebrook Research Station, Wareham, Dorset UK BH20 5AS, United Kingdom and [‡]Department of Mathematical Sciences, Lade Section, Norwegian University of Science and Technology, 7034 Trondheim, Norway*

Manuscript received August 27, 1997
Accepted for publication March 31, 1998

ABSTRACT

A model of the migration pattern in a metapopulation of sea beet (*Beta vulgaris* L. ssp. *maritima*), based on the continuous distributions of seed and pollen movements, is fitted to gene frequency data at 12 isozyme and RFLP loci by maximum likelihood by using an approximation of the simultaneous equilibrium distribution of the gene frequencies generated by the underlying multivariate stochastic process of genetic drift in the population. Several alternative restrictions of the general model are fitted to the data, including the island model, a model of complete isolation, and a model in which the seed and pollen dispersal variances are equal. Several likelihood ratio tests between these alternatives are performed, and median bias in the estimated parameters is corrected by using parametric bootstrapping. To assess the fit of the selected model, the predicted covariances are compared with covariances computed from the data directly. The dependency of estimated parameters on the ratio between effective and absolute subpopulation sizes, which is treated as a known parameter in the analysis, is also examined. Finally, we note that the data also appear to contain some information about this ratio.

LEVELS of gene flow can be estimated from different kinds of data, either directly by different forms of capture-recapture methods or indirectly from geographic genetic differentiation generated by local genetic drift. Even though all information in genetic data on which indirect methods are based lies in the variances and covariances of the gene frequencies only, indirect methods are potentially very useful because data at several loci represent independent realizations of the same underlying process. If appropriate assumptions are built into the analysis, indirect methods may also potentially produce gene flow estimates reflecting average levels several years into the past (Slatkin 1985), thus being more relevant in an evolutionary context.

Although some theoretical results that allow inferences to be drawn about the pattern of migration from observations of genetic geographic differentiation exist, these theoretical models are often highly idealized. In the island model of migration, for example, it is assumed that all migrations occur via some large outside world populations (Wright 1943). If we believe this assumption, the effective number of migrants that enter each subpopulation each generation, N_em, can be calculated from the estimated amount of genetic differentiation between the subpopulations. In Kimura and Weiss' (1964) analysis of decrease in genetic correlation in infinite stepping stone models, migration also occurs between neighboring subpopulations. It is also assumed that the subpopulations constitute completely panmictic, discrete, regularly spaced units. In most natural populations, however, individuals are more or less continuously distributed across space in varying densities. In most plant species, dispersal of genes, through seed and pollen, also occur over distances of almost any length. The distributions of these seed and pollen displacements are often highly leptokurtic.

To obtain reliable and useful estimates of levels of gene flow, it seems necessary to incorporate more realistic assumptions into the analysis and to take into account the geographic structure of the population under study, variation in local subpopulation sizes, and the form of dispersal. In this article, we analyze a set of gene frequency data (Raybould *et al.* 1996) from a metapopulation of sea beet (*Beta vulgaris* L. ssp. *maritima*) by using and further developing methods from Tufto *et al.* (1996). In this approach, the likelihood of different migration matrices, given the data, is computed using an approximation of the simultaneous stationary distribution of the gene frequencies generated by the underlying multivariate stochastic process of genetic drift in the subpopulations under consideration. After presenting the data and a brief summary of the basic approach, we introduce a general model of migration patterns in plant populations in which the migration matrix is related to the continuous dispersal probability distributions of seed and pollen movements. Two such dispersal distributions applying specifically to the data used are then given. The parameters of these distributions are then estimated from the gene frequency data by maximum likelihood.

Corresponding author: Jarle Tufto, Department of Mathematical Sciences, Lade Section, Norwegian University of Science and Technology, 7034 Trondheim, Norway. E-mail: jarlet@math.ntnu.no

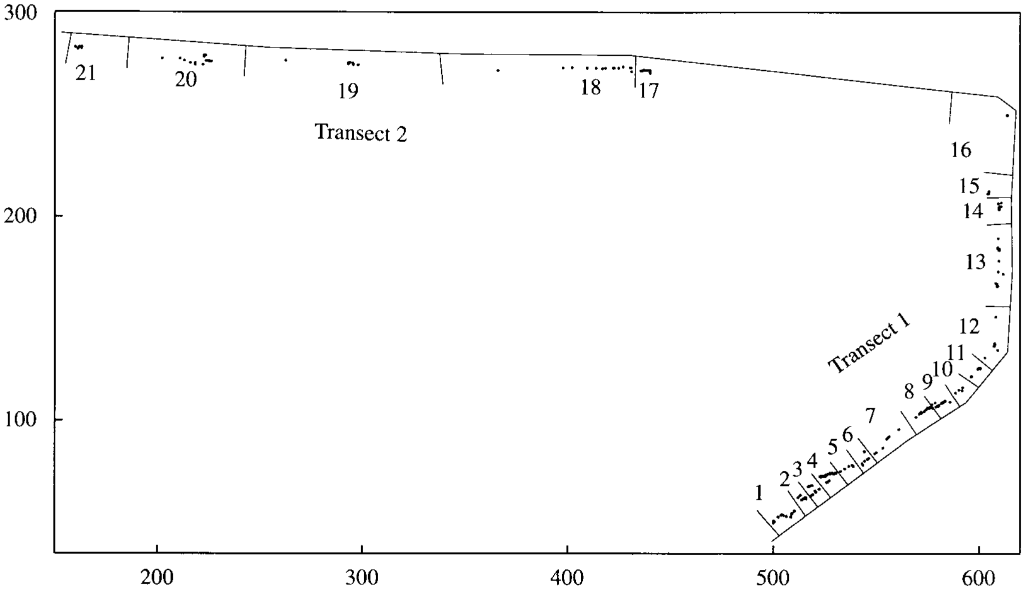

Figure 1.—The geographic location of the 220 sea beet plants (represented by the dots) divided into subpopulations 1–21. The coordinate system is in units of meters. In the result and discussion sections, subpopulations 1–16 and 17–21 will be referred to as transects 1 and 2, respectively.

Because only an approximate likelihood is computed and the data are limited, we can expect estimates of the different parameters to be biased. By making extensive use of bootstrapping methods, however, these biases can be corrected to a great extent. Using parametric bootstrapping from several restrictions of the general model of the migration pattern, likelihood ratio tests between some alternative restricted models, including the island model, are also performed. The fit of the selected model is then inspected by comparing predicted and observed covariances. Finally, we examine the dependency of the estimated parameters on the assumed ratio between effective and absolute subpopulation sizes. We also note that the data, which are purely nontemporal, appear to contain some information about local effective subpopulation sizes, possibly because time in the underlying process is discrete.

THE DATA

The data consist of allele frequencies of 12 isozyme and RFLP loci from 220 sea beet plants at Furzey Island, Poole Harbour, UK, that were previously analyzed in Raybould *et al.* (1996). The coordinates of all sampled plants in space and the subdivision used to construct subpopulations are shown in Figure 1.

Because almost all plants seen in the study area were sampled, the absolute subpopulation sizes, which are part of the model input, could be counted directly. This also means that the gene frequencies in subpopulations could be determined without any sampling error.

The model also involves the effective sizes of each of the subpopulations, which are generally much harder to determine. The effective size N_e of a population will generally equal the absolute population size N, multiplied by some factor that depends on demographic parameters. Sea beet is a self-incompatible, iteroparous perennial with a life span of at least 3 years in cultivation (Bruun *et al.* 1995; Van Dijk *et al.* 1997). Personal observations suggest that plants in the Dorset populations live for at least 5 years. Although sea beet is gynodioecious in France (Boutin *et al.* 1988), the Dorset plants are all hermaphrodite. According to Nunney (1993), there is a general tendency for N_e to approach $N/2$ as generation length and generation overlap increases. We have therefore used this limiting value for the effective size of each subpopulation, although estimates of this ratio are known to vary widely from ~0.2 to 0.8 (Frankham 1995).

COMPUTING THE LIKELIHOOD

Tufto *et al.* (1996) present a general approach for estimating the matrix **M** of migration rates between a set of populations, $i = 1,2,...,n$, from geographic neutral genetic variation, in part using ideas in Courgeau (1974) and Felsenstein (1982). The model is based on the underlying multivariate stochastic process

$$\mathbf{p}_{t+1} = \mathbf{M}\mathbf{p}_t + \mathbf{e}_t \qquad (1)$$

of genetic drift in the vector $\mathbf{p}_t = (p_{1,t},...,p_{n,t})$ of the gene frequencies in each subpopulation. The elements of the column vector $\mathbf{e}_t$ representing genetic drift have variances equal to $p_{i,t}(1 - p_{i,t})/2N_i$, where N_i is the effective size of subpopulation i. For this process, the exact covariance matrix **C** of the stationary distribution of the standardized gene frequencies can be found by solving the matrix equation

$$\mathbf{C} = \mathbf{M}\mathbf{C}\mathbf{M}^T + \mathbf{E}, \qquad (2)$$

where

$$e_{ij} = \begin{cases} \dfrac{1 - c_{ii}}{2N_i} & \text{for } i = j \\ 0 & \text{for } i \neq j. \end{cases} \tag{3}$$

using numerical techniques (Tufto *et al.* 1996; appendix). This can be done for migration matrices of any form, for systems with a moderate number of subpopulations.

The basic idea of the approach is to assume that the migration matrix is of a certain form, *e.g.*, stepping stone like, then compute the covariance matrix from the migration matrix and the effective population sizes, and then use the multivariate normal distribution to repeatedly compute the probability of the observed gene frequencies for different parameter values until the likelihood is maximized. With gene frequency data on several independent loci, the total likelihood is equal to the product of the multinormal probabilities calculated for each loci.

Conditioning on $\bar{p}$: The model initially also involves an $(n + 1)$th large "outside world" population with gene frequency remaining constant over time and equal to q. To eliminate this nuisance parameter from the model, the distribution of the gene frequencies, conditioned on a sufficient statistic for q, have to be considered. In the case of multivariate normality, which would hold if the fluctuations around the equilibrium gene frequency are small, the weighted mean observed gene frequency $\bar{p} = \Sigma_i w_i p_i$ is known to be sufficient for q, provided that the w_i's are chosen to minimize the variance of $\bar{p}$. The sampling distribution of the model, conditioned on $\bar{p}$, is then, by definition, independent of q. The major difficulty of the approach is to find a good approximation of the covariance matrix of this conditional distribution, more generally, when multivariate normality does not hold (see appendix). The approximation of the conditional covariances do not have to be completely accurate, however. As we shall see, as long as we are able to simulate bootstrap data from the correct sample distribution, biases, which in part may be caused by approximations involved in the computation of the likelihood, can be corrected to a great extent.

MIGRATION MATRICES IN PLANT POPULATIONS

General considerations: Following Bodmer and Cavalli-Sforza (1968), each element m_{ij} of the matrix $\mathbf{M}$ in (1), the backward migration matrix, is defined as the conditional probability that a gene originates from population j, given that it dispersed to population i. The purpose of this section is to define a biologically realistic model for the migration pattern, which is essentially a specification of $\mathbf{M}$ as a function of the basic parameters involved.

The migration probabilities will generally depend on the relative number of individuals in each subpopulation because large subpopulations will contribute with a greater proportion of genes to subsequent generations. It must also be remembered that gene movements in plant populations occur in either one or two stages: through seed dispersal only or through both pollen and seed dispersal, possibly via some third population.

Let $\mathbf{M}^{(s)}$ and $\mathbf{M}^{(p)}$ be the backward migration matrices for seed and pollen dispersal, *i.e.*, $m_{ij}^{(s)}$ and $m_{ij}^{(p)}$ denote the conditional probabilities that a seed or a pollen grain, respectively, originates from subpopulation j, given that it dispersed to subpopulation i. Consider the event that dispersal occurs through both pollen and seed via some third population k. The conditional total probability that the gene originates from j is then $\Sigma_k m_{kj}^{(p)}\ m_{ik}^{(s)}$. Given the other event that the gene dispersed through a seed only, the probability that it originates from j will be $m_{ij}^{(s)}$. Because all zygotes in any generation are formed by the union of a pollen grain and a seed, these two events both have probability ½, implying that $m_{ij} = \frac{1}{2} \Sigma_k m_{kj}^{(p)}\ m_{ik}^{(s)} + \frac{1}{2} m_{ij}^{(s)}$ or, in matrix notation,

$$\mathbf{M} = \tfrac{1}{2}\mathbf{M}^{(s)}\,(\mathbf{M}^{(p)} + \mathbf{I}). \tag{4}$$

For each form of dispersal, $\nu \in \{s,p\}$, let the forward matrix $\tilde{\mathbf{M}}^{(\nu)}$ with elements $\tilde{m}_{ij}^{(\nu)}$ represent the probabilities that a seed or a pollen grain disperses to subpopulation i, given that it originates from subpopulation j. As noted by Bodmer and Cavalli-Sforza (1968), the relationship between the elements of the forward and backward migration matrices $\tilde{\mathbf{M}}^{(\nu)}$ and $\mathbf{M}^{(\nu)}$ is

$$m_{ij}^{(\nu)} = \frac{\tilde{m}_{ij}^{(\nu)}\mathrm{N}_j}{\Sigma_j \tilde{m}_{ij}^{(\nu)} N_j}. \tag{5}$$

For self-incompatible species, the N_j's in (5) for $\nu = p$ should be replaced by $N_j - 1$ for $j = i$.

Let $f_{\mathbf{z}}^{(\nu)}(\mathbf{z})$, $\nu \in \{s,p\}$, represent the distribution of the random displacement $\mathbf{Z}$ of seed or pollen occurring during dispersal in the relevant spatial dimension(s). In general, for a seed or pollen originating from some point in subpopulation j, the probability that it disperses to some point in another population i will be

$$\tilde{m}_{ij}^{(\nu)} = c_i E f_{\mathbf{z}}^{(\nu)}(\mathbf{Z}_j - \mathbf{Z}_i), \tag{6}$$

where the expectation is taken over points $\mathbf{Z}_j$ and $\mathbf{Z}_i$ in subpopulations i and j. The constant c_i represents the effects of local environmental conditions on the success of the dispersing seed or pollen grain in the receiving subpopulation i. The local environment c_i may, for example, involve local population density and the spatial size of the receiving subpopulation. These effects, however, do not influence the resulting backward migration probabilities, which is seen when (6) is substituted into (5).

For neighboring subpopulations and, at least, for dispersal within subpopulation, the expectation in (6) must be computed numerically. For the sea beet data used in this article, for which the positions of all individual plants were available, this is straightforward to do. When the distance between two populations is large, the expectation can be computed using the first-order approximation $f_{\mathbf{z}}^{(v)}(E\mathbf{Z}_j - E\mathbf{Z}_j)$.

Seed and pollen dispersal distribution in sea beet: In sea beet, pollen is wind dispersed, and the resulting dispersal distributions are thus two dimensional. Tufto *et al.* (1997) discuss several plausible forms that $f_{\mathbf{z}}^{(\mathrm{p})}(\mathbf{z})$ may take in such cases, arising from different sets of assumptions about the underlying physical movement process. If we assume that wind conditions change slowly over time, that pollen are released from the anthers when the wind velocity exceeds a certain threshold, that the movement of each individual pollen grain is a diffusion in three dimensions, and that there is no wind directionality in the long run, the various parameters collapse into a single parameter, λ_{p}, and the dispersal distribution function takes the form

$$f_{X,Y}^{(\mathrm{p})}(x,y) \propto \frac{1}{r^{3/2}}e^{-\lambda_{\mathrm{p}}r}, \qquad (7)$$

where $1/2\lambda_{\mathrm{p}}$ is the expectation of $R = \sqrt{X^2 + Y^2}$ (for details see Tufto *et al.* 1997). Compared with several alternative models, an extension of (7) gave the best fit to experimental data on airborne pollen dispersal in meadow fescue (*Festuca pratensis* L.) using genetic markers (Nurminiemi *et al.* 1997).

The populations of sea beet considered in this article were located along the shoreline, and seed dispersal was known to occur mostly by the help of tidal currents moving back and forth along the shoreline in approximately one dimension, *w*. If we assume that seeds are equally likely to be released into the water at any point in time, so that dispersals in both directions have equal probability, and also assume that the probability of deposition in a small length interval w, $w + dw$ is $\lambda_{\mathrm{s}}dw$, then the random displacement W of each seed will follow the Laplace distribution

$$f_W^{(\mathrm{s})}(w) = \frac{1}{2}\lambda_{\mathrm{s}}e^{-{}_{\mathrm{s}}|w|}, \qquad (8)$$

also known as the double exponential distribution. This distribution also arises by assuming that the time to deposition is exponential and that the movement path $W(t)$ is a Brownian motion without drift (Kendall *et al.* 1983, p. 191).

Note that the seed and pollen dispersal distributions (7) and (8) differ greatly in their degree of kurtosis, *i.e.*, to what extent probability is concentrated around the mean and in the tails of the distribution. This difference is in part a result of the assumptions that pollen dispersal occurs in two dimensions whereas seeds only disperse in approximately one dimension (along the shoreline).

We also assume that the backward probabilities of migration from the outside world are equal to u for all subpopulations. This may be a questionable assumption in part because the subpopulations differ in their density and in their distance from the outside world. However, if these backward migration rates are small, this assumption may not be very critical. The parameter u can also be thought of as a mutation rate, in which case, it would be the same for all subpopulations.

BOOTSTRAPPING METHODS

Simulation procedure: Several strategies can be used to simulate the distribution of the data. One strategy would be to reconstruct the distribution from the data itself by resampling gene frequency vectors with equal probability from each of the observed loci. This, however, would not take advantage of the knowledge we have about the process generating the data. A better strategy is to stimulate the process directly by iterating (1) for a sufficient number of generations until the equilibrium distribution is attained. It must be kept in mind, however, that it is the distribution conditional on the observed gene frequency means at each locus that is relevant. This conditional distribution can be generated approximately by first simulating, say, 500 generations, then keeping track of the gene frequency mean for each iteration of the process, and stopping the iterations just after (or before) the gene frequency mean passes the target mean. This simulation technique eliminates the extra variance in the parameter estimates that would arise if the unconditional distribution, which very likely would contain nonpolymorphic data, was used.

Bias correction: The sampling distribution of an estimator $\bar{\theta}$ usually has a peak near or at the unknown parameter θ of interest. We say that an estimator $\hat{\theta}$ is mean unbiased if $E\hat{\theta} = \theta$. By simulating the distribution of the estimator, provided that the expectation of the estimator exists, one can always compute a less biased estimate $\bar{\theta}$ by subtracting the estimated amount of bias from the original first estimate (Efron and Tibshirani 1993, p. 138).

This procedure, however, is not invariant with respect to transformations of the parameter, and in cases where the distribution of the estimator is very skewed, it can often make more sense to construct a so-called median bias-corrected estimate $\bar{\theta}$ (Cabrera and Watson 1997). This is done by estimating the median, $M\hat{\theta}$, of the estimates as a function $g(\theta) = M\hat{\theta}$ of the parameter. The improved estimate is then the value of the unknown parameter for which the median of the bootstrap estimates matches the observed estimate, *i.e.*, $\bar{\theta} = g^{-1}(\hat{\theta})$ (see Figures 2 and 4). As shown by Cabrera and Watson (1997), provided that $g(\theta)$ is a monotonic function, it then follows that the median bias-corrected estimate $\bar{\theta}$ itself is median unbiased because $M\bar{\theta} = M(g^{-1}(\hat{\theta})) = g^{-1}(M(\hat{\theta})) = g^{-1}(g(\theta)) = \theta$.

Likelihood ratio tests: Several tests between different hypothesis about the unknown migration pattern will be of interest. In general, in tests of a model H_0 against some other model H, the ratio Λ of the maximum likelihood under H and H_0 is usually used as a test statistic. If there is a small probability, given that H_0 is true, that Λ is larger than the value of Λ computed from the observed data, Λ^*, H_0 can be rejected in favor of H. This probability, $P(\Lambda > \Lambda^*|H_0)$, will also generally depend on the true value of the nuisance parameters θ_0 under H_0. A reasonable choice of θ_0 to simulate from is the value of θ_0 that is in best agreement with the observed data, which in our case is the median bias-corrected estimate $\bar{\theta}$ obtained by fitting H_0. This choice of θ_0 will in many (Efron and Tibshirani 1993, p. 232) but not all cases make the test conservative; *i.e.*, the test is less likely to falsely reject the null hypothesis.

RESULTS

Model selection: Four different restrictions of the general model of the migration pattern were considered. We first fitted the model under the constraint that the variances of the seed and pollen dispersal displacements, $\mathrm{Var}(X) = \mathrm{Var}(Y) = 3/8\lambda_p^2$, and $\mathrm{Var}(W) = 2/\lambda_s^2$, are both equal, using u as the other free parameter, *i.e.*,

$$\frac{3}{8\lambda_p^2} = \frac{2}{\lambda_s^2} = \sigma^2 \geq 0, \quad u \geq 0. \tag{9}$$

The uncorrected point estimates of the seed and pollen dispersal standard deviation and the long-range migration rate obtained under this model were $\hat{\sigma} = 60.4m$ and $\hat{u} = 0.062$, and the maximum likelihood was 122.57.

We then tested this model against the alternative extended model using both λ_s and λ_p as free parameters in addition to u:

$$\frac{3}{8\lambda_p^2} \neq \frac{2}{\lambda_s^2}, \quad u \geq 0. \tag{10}$$

This gave only a very slight increase amounting to 0.05 in the maximum likelihood, which is clearly statistically nonsignificant if relying on asymptotic theory as a rough approximation. Two hypotheses nested within (9) were then considered, the first defined by the constraint

$$\sigma = 0, \quad u \geq 0, \tag{11}$$

corresponding to the classical island model (Crow and Kimura 1970). Note that we have assumed that the rate of migration, and not the effective number of migrants, N_iu, is constant. The bias-corrected estimate of u obtained by fitting (11) was $\bar{u} = 0.277$ (Figure 2). Simulating the distribution of Λ at this point in the parameter space, we could reject the island model in favor of model (9) (Figure 3).

The second hypothesis nested within (9) is defined by the constraints

$$\sigma \geq 0, \quad u = 0, \tag{12}$$

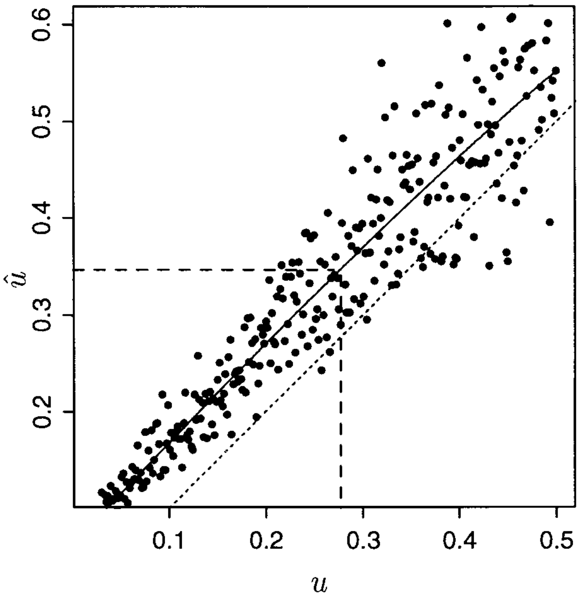

Figure 2.—The sampling distribution of $\hat{u}$ for different values of the long-range migration rate u under the island model. The solid line is the median of $\hat{u}$ as a function of u, estimated using minimum absolute residual regression (function l1fit in S-Plus) with a third-order polynomial as regressor. The value of the improved, bias-corrected estimate $\bar{u}$ corresponding to the observed value of $\hat{u}$ is indicated by the dashed line on the graph.

corresponding to the hypothesis that the metapopulation under study is completely isolated from the outside world. The distribution of the data (and estimators and test statistics) under this null hypothesis can be simulated approximately by replacing $u = 0$ in the iterations of (1) with some small positive value, say, 10^{-4}. This is necessary to avoid letting the iterations go into an infinite loop, but if a sufficiently small number is used, the approximation will have the necessary accuracy.

Fitting (12) to the data gave an uncorrected estimate

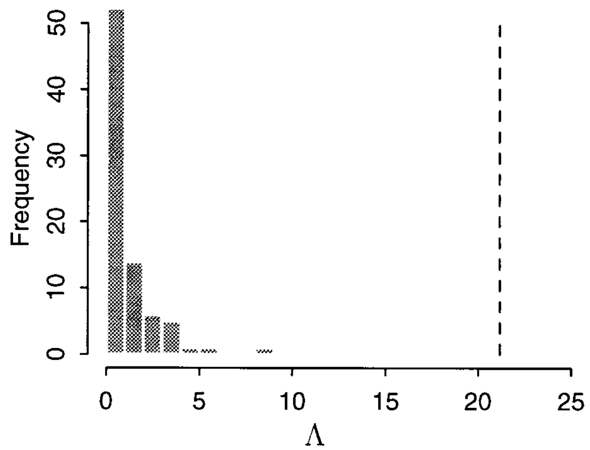

Figure 3.—Histogram of 100 simulations of the log likelihood ratio, Λ, between models (11) and (9). The observed value of Λ is indicated by the dashed line.

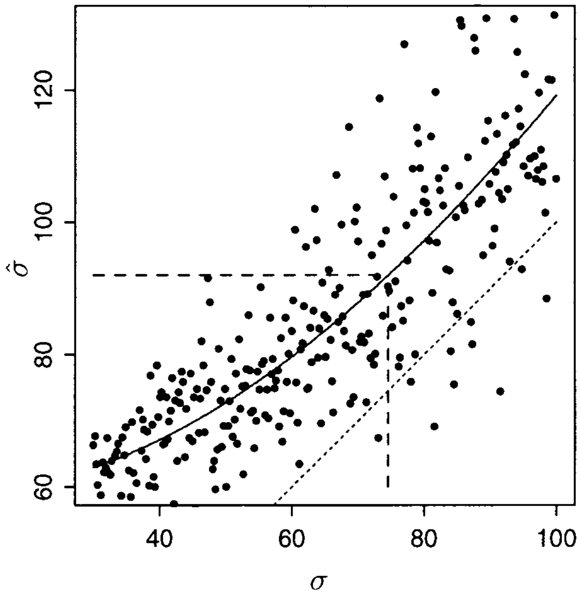

Figure 4.—The sampling distribution of $\hat{\sigma}$ for different values of the seed and pollen dispersal standard deviation σ under the complete isolation model (12). For further explanation, see Figure 2.

of the seed and pollen dispersal standard deviation of $\hat{\sigma} = 92.0\ m$ and a maximum likelihood of 116.8. Simulating the distribution of $\hat{\sigma}$ for different values of σ (Figure 4), we obtained a bias-corrected estimate, $\overline{\sigma} = 74.56\ m$. On the basis of the simulation of the distribution of the likelihood ratio between models (9) and (12), using the bias-corrected estimate $\overline{\sigma}$ for σ, we could not reject the hypothesis of complete isolation (12). A histogram of these log likelihood ratios are shown in Figure 5.

Observed and predicted covariances: To assess the fit of the selected model, we will compare the (conditional) standardized covariances predicted by the model with those computed from the data directly. These are shown in Figure 6. Some interesting points can be noted. First, there appears to be reasonably good agreement between the model predictions and the observations. The main level of isolation is between the two transects (see Figure 1). This can be seen in the predicted values as "steps" in the covariances between subpopulations 16 and 17. Within the less dense transect 2 (subpopulations 17–21), there is also a steeper decline in the covariances with increasing geographic distance than within transect 1, where there is only a slight effect of isolation by distance. There also appears to be reasonably good agreement between predicted and observed variances, in that relatively small and/or isolated subpopulations, *e.g.*, subpopulations 15 and 16, have larger predicted and observed variances.

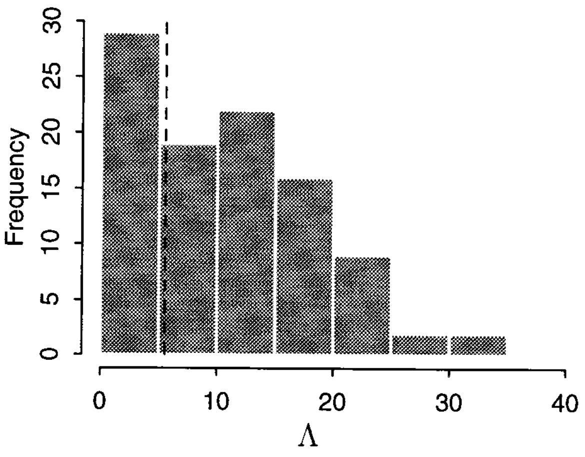

Figure 5.—Histogram of 100 simulations of the log likelihood ratio, Λ, between models (12) and (9). The observed value of Λ is indicated by the dashed line.

The estimated migration matrix: The migration matrix computed at the bias-corrected estimate of σ is shown in Figure 7. On average, there is ~70% migration of genes from neighboring populations. Large, isolated, dense subpopulations, such as subpopulation number 20, receive a smaller proportion from neighboring subpopulations, whereas very small subpopulations (*e.g.*, number 10 consisting of 4 individuals) receive a larger proportion of genes from its neighbors than from itself. It is also apparent in several subplots that large subpopulations (*e.g.*, subpopulation number 4 consisting of 29 individuals) also make a large contribution to distant subpopulations because of their size.

Having estimated the variances in the seed and pollen dispersal displacements under the assumption that these are equal, the total variance of the gene displacements, measured along the shoreline, if we assume independence between X and W, becomes Var(W) + ½Var(X) (Crawford 1984), which gives a gene displacement standard deviation equal to 91.31 m for model (12).

Dependencies on N_e/N: Because the data are nontemporal, we do not expect much information about the ratio between the effective and absolute population size of each subpopulation, N_e/N, to be available in the data, and this ratio has, therefore, so far been treated as a known parameter in the analysis described above. Because our choice of $N_e/N = ½$ is clearly rather arbitrary, we will, however, look at how our estimate of σ under the selected model (12) depends on the choice of N_e/N. The maximum likelihood estimate of σ obtained by fitting the model for each of several different N_e/N values is shown in Figure 8.

First note that on a log–log scale, for values of N_e/N larger than ~0.2, there appears to be a linear relationship with $\hat{\sigma}$, with a slope of ~−½; *i.e.*, the estimate of the dispersal standard deviation, σ, is proportional to $(N_e/N)^{-1/2}$ or, equivalently, the dispersal variance is inversely proportional to the effective population sizes. This is analogous to the situation under the island model, where only the product between the effective size and the rate of migration into each subpopulation,

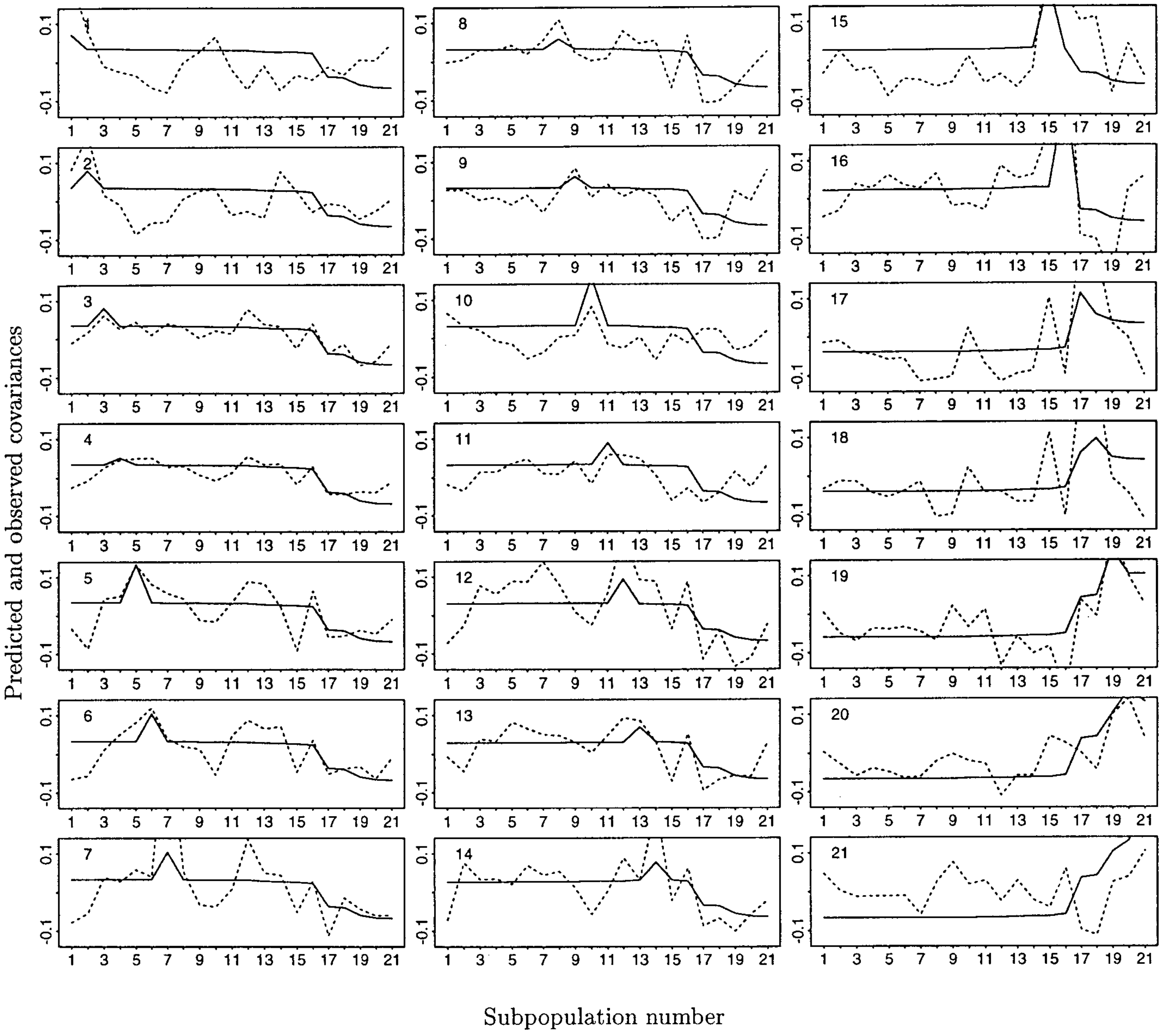

Figure 6.—Observed covariances (dotted lines) and covariances predicted by the selected model (12) (solid lines). Each subplot represents single rows in the respective matrices. The first plot in the third column, *e.g.*, is the predicted and observed covariances of subpopulation number 15 with all other subpopulations (including the variance with itself).

or the "effective number of migrants," is possible to estimate.

Also note that the estimate of σ appears to tend to infinity when N_e/N becomes smaller than ~0.15. The estimate of σ is then about the same order of magnitude as the size of the study area. Any further increase in σ then has little influence on the migration matrix, and hence the likelihood, which means that the maximum likelihood estimate of σ tends to infinity.

We expected the data to contain almost no information about N_e/N, making the maximum likelihood almost independent of this parameter. To verify this, we plotted the maximum likelihood obtained by fitting the model for each of the different N_e/N values. This is shown in Figure 9. Surprisingly, the likelihood is strongly dependent on N_e/N and has its maximum at ~0.43.

DISCUSSION

We have shown how a quite realistic model based on the underlying genetic multivariate drift process in a subdivided population can be fitted to gene frequency data by approximate maximum likelihood methods. The model uses information known about absolute and effective population sizes, as well as geographic distances within and between subpopulations. Assumptions about the underlying mechanisms of seed and

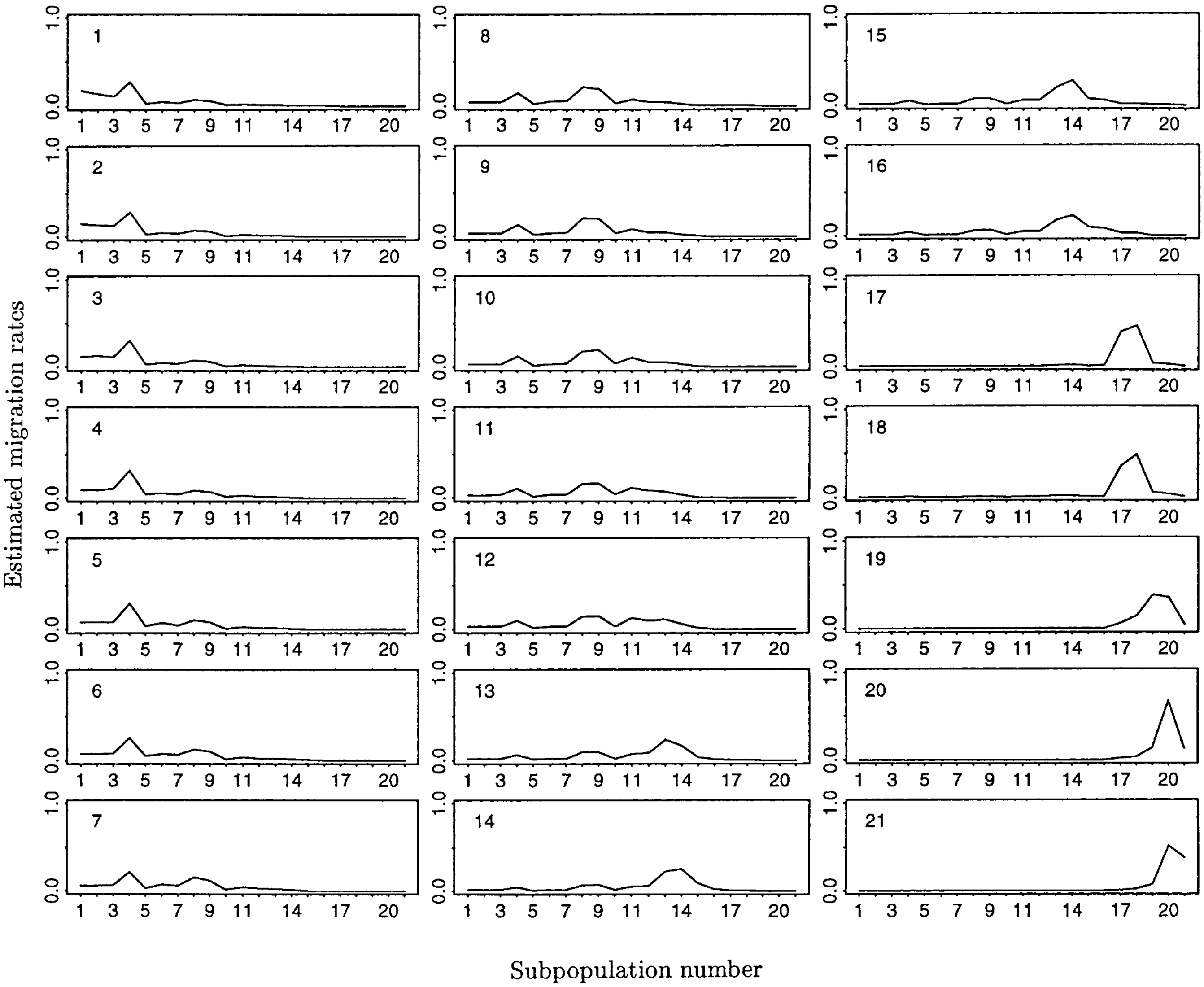

Figure 7.—The estimated migration matrix under the model of complete isolation computed at the bias-corrected estimate of σ. Each subplot represents a single row in the backward migration matrix.

pollen movement are also incorporated. A data set consisting of allele frequencies at 12 isozyme and RFLP loci in 21 subpopulations of sea beet was used, and inferences about at least one parameter, the seed and pollen displacement variance, could be made. We were also able to reject the island model in favor of our more general model. It could also be concluded that no significant evidence for migration from the outside world or for mutation is present in the data.

There appears to be quite good agreement between the observed and predicted covariances. However, the fact that the observed covariances are dependent, also given that the selected model is true, makes comparison with the covariances predicted by the fitted model dangerous. If we look at Figure 6, there appear to be some discrepancies between the observed and predicted correlations with subpopulation number 15, in that subpopulation 15 appears to be more similar to subpopulations 16–21 and less similar to subpopulations 1–14 than predicted by the fitted model. This may suggest that some isolating barrier is present between subpopulations 14 and 15. Because the correlations are dependent, however, this discrepancy may very well be an effect of chance, and it would be desirable to do some more formal test of goodness of fit. As suggested by Felsenstein (1982), this can be done by testing the selected model against a "full" model in which all the $n(n-1)/2$ covariances of the $n-1$ contrasts are used as free parameters. Such a test would require data on a large number of loci, however, partly because the number of parameters under the full model is very large. The observed covariance matrix will also be singular if the number of loci is less than $n-1$, which excludes the possibility of doing this test in most cases.

If the degree of kurtosis of the gene displacements has an effect on the genetic structure, we could hope

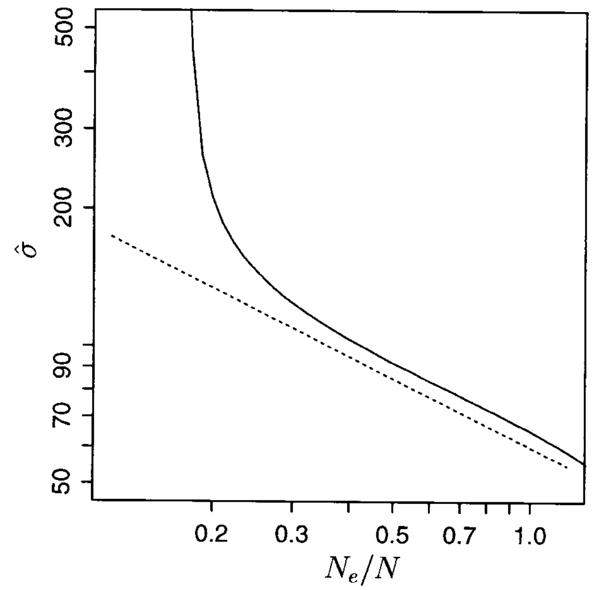

Figure 8.—The (uncorrected) maximum likelihood estimate of σ for different values of N_e/N. The dotted reference line has a slope equal to $-\frac{1}{2}$.

to be able to obtain separate estimates of the seed and pollen dispersal variances because these distributions, (7) and (8), differ greatly in their degree of kurtosis. However, the fact that the extended model (10) gave almost no increase in the likelihood suggests that the data contain very little information about this.

In a previous, more traditional analysis of the data than the one used here, pairwise genetic distances were regressed against geographic distances (Raybould *et al.* 1996). These authors also found a "step" in the genetic distances between the two transects by including dummy variables representing transect membership in the regression, and they concluded that this suggested continuous extinction and recolonization of the subpopulations. However, the analysis used here shows that the "step" in the covariances between subpopulations 16 and 17 can be fully explained by the geographic location of the subpopulations only (Figure 1) and the assumed pattern of migration.

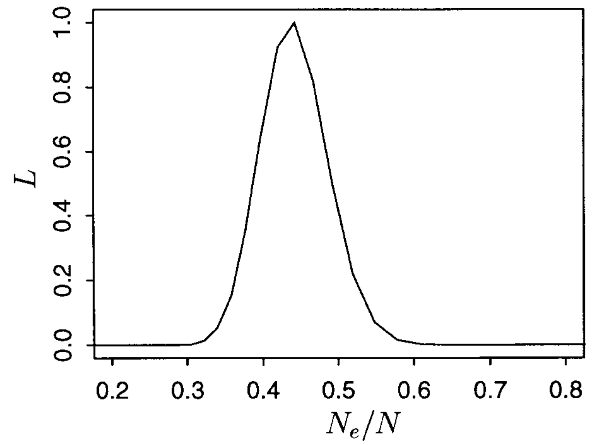

Figure 9.—The maximum likelihood as a function of N_e/N.

We still need to be concerned about the assumption that population sizes remain constant over time, however. If there are large autocorrelated fluctuations in local population sizes, this is likely to constantly keep the gene frequencies away from their equilibrium distribution if the number of generations necessary to reach equilibrium is large. It is, however, only relevant to consider the rate of convergence of the covariances of the contrast, and these will converge much faster than the covariance matrix of the entire distribution (on the order of only several generations; Harpending and Ward 1982, p. 245).

It is also very interesting that there appears to be information in the data about the ratio between effective and absolute subpopulation sizes. This is quite suprising when considering the fact that the data are purely nontemporal. It is well known that in continuous time diffusion approximations of, *e.g.*, the island model, the rate of migration and the effective population size of each island are completely confounded. However, in our study population and in the model, generations are discrete. The effective subpopulation sizes are also quite small, between 1 and 14.5 individuals, and the rates of migration between neighboring populations are relatively high, which suggest that there will be quite large changes in the equilibrium covariances at different points in the life cycle. The magnitude of these changes depends on N_e/N. Even though the maximum likelihood estimate of $N_e/N = 0.43$ obtained seems plausible, it is not clear how reliable this estimate is, however. One potential problem is that the model has nonoverlapping generations, whereas generations overlap in the study population, possibly also with age-dependent mortalities and fecundities. It is also not entirely clear if the point in time in the life cycle of the predicted covariances coincides with the point in time at which the actual sampling occurred.

There has been some concern in the literature that direct and indirect methods have produced very different estimates of gene flow. There is a general tendency for indirect estimates to be much larger than estimates obtained by direct methods (*e.g.*, Koenig *et al.* 1996). According to these authors, the discrepancy results from systematic biases in direct estimates caused by finite study areas. We note here that our estimate of seed and pollen dispersal standard deviations of 74.56 *m* are comparable to estimates from direct studies of seed dispersal that range between 27 and 100 m (Slatkin 1985) and to estimates of pollen dispersal of 30 m (Nurminiemi *et al.* 1997).

Availability of software: The S-Plus and C code used to do this analysis and accompanying documentation is available on the internet at http://www.math.ntnu.no/~jarlet/migration/

We thank Ralph Clarke for helpful discussions. This study was financially supported by the Norwegian Research Council and the Norwegian Institute for Nature Research A.F.R. also received financial support from the Department of Environment's Genetically Modified Organisms Research Program.

LITERATURE CITED

Bodmer, W. F., and L. L. Cavalli-Sforza, 1968 A migration matrix model for the study of random genetic drift. Genetics **59:** 565–592.

Boutin, V., R. Jean, M. Valero and P. Vernet, 1989 Gynodioecy in *Beta maritima.* Acta Oecologia **9:** 61–66.

Bruun, L., A. Haldrup, S. G. Petersen, L. Frese, T. de Bock *et al.*, 1995 Self-incompatibility reactions in wild species of the genus beta and their relation to taxonomical classification and geographical origin. Genet. Resourc. Crop Evol. **42:** 293–301.

Cabrera, J., and G. S. Watson, 1997 Simulation methods for mean and median bias reduction in parametric estimation. J. Stat. Plan. Infer. **57:** 143–152.

Courgeau, D., 1974 Migration, pp. 351–387, in *The Genetic Structure of Populations*, edited by A. Jacquard. Springer-Verlag, Berlin.

Crawford, T. J., 1984 What is a population? pp. 135–173, in *Evolutionary Ecology*, edited by B. Shorrocks. Blackwell Scientific Publications, Oxford.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory.* Harper & Row, New York.

Efron, B., and R. J. Tibshirani, 1993 *An Introduction to the Bootstrap.* Chapman & Hall, London.

Felsenstein, J., 1982 How can we infer geography and history from gene frequencies? J. Theor. Biol. **96:** 9–20.

Frankham, R., 1995 Conservation genetics. Annu. Rev. Genetics **29:** 305–327.

Harpending, H. C., and R. H. Ward, 1982 Chemical systematics and human populations, pp. 213–256, in *Biochemical Aspects of Evolutionary Biology*, edited by M. H. Nitecki. University of Chicago Press, Chicago.

Kendall, M., A. Stuart and J. K. Ord, 1983 *Kendall's Advanced Theory of Statistics*, Vol. 1. Charles Griffin and Company Ltd., London.

Kimura, M., and G. H. Weiss, 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics **49:** 561–576.

Koenig, W. D., D. V. Vuren and P. N. Hooge, 1996 Detectability, philopatry, and the distribution of dispersal distances in vertebrates. Trends Ecol. Evol. **11:** 514–517.

Nunney, L., 1993 The influence of mating system and overlapping generations on effective population size. Evolution **47:** 1329–1341.

Nurminiemi, M., J. Tufto, N.-O. Nilsson and O.-A. Rognli, 1997 Spatial models of pollen dispersal in the forage grass meadow fescue. Evol. Ecol. **12:** 487–502.

Raybould, A. F., J. Goudet, R. J. Mogg, C. J. Gliddon and A. J. Gray, 1996 Genetic structure of a linear population of *Beta vulgaris* ssp. *maritima* (sea beet) revealed by isozyme and RFLP analysis. Heredity **76:** 111–117.

Slatkin, M., 1985 Gene flow in natural populations. Ann. Rev. Ecol. Syst. **16:** 393–430.

Tufto, J., S. Engen and K. Hindar, 1996 Inferring patterns of migration from gene frequencies under equilibrium conditions. Genetics **144:** 1911–1921.

Tufto, J., S. Engen and K. Hindar, 1997 Stochastic dispersal processes in plant populations. Theor. Pop. Biol. **52:** 16–26.

Van Dijk, H., P. Boundry, H. McCombie and P. Vernet, 1997 Flowering time in wild beet (*Beta vulgaris* ssp. *maritima*) along a latitudinal cline. Acta Oecologia **18:** 47–60.

Wright, S., 1943 Isolation by distance. Genetics **28:** 114–138.

Communicating editor: B. S. Weir

APPENDIX: APPROXIMATIONS OF THE CONDITIONAL COVARIANCE MATRIX

In Tufto *et al.* (1996), an approximation based on (2) directly was shown to break down as the fluctuations around q become large. On the basis of the idea that the amount of genetic drift in the neighborhood of the observed mean gene frequency is most important in determining $\mathbf{C}_{y|\bar{p}}$, and in an attempt to also incorporate the fact that the genetic drift depends on the gene frequencies, Tufto *et al.* (1996) developed a second approximation of $\mathbf{C}_{y|\bar{p}}$ by introducing a linear matrix transformation of the state vector in the recurrence equation (1), changing the dynamics of the process and forcing it to move within the subspace where $\sum_i w_i p_i = \bar{p}$ only. This was done by letting

$$\mathbf{p}_{t+1} = \mathbf{D}(\mathbf{M}\mathbf{p}_t = \mathbf{e}_t). \qquad \text{(A1)}$$

This leads to a matrix equation in $\mathbf{C}$ (Tufto *et al.* 1996, Eq. 18) similar to (2). As long as there is a limited amount of differentiation, maximum likelihood estimates of the parameters of a finite stepping stone model based on this approximation using simulated data were shown to be reasonably unbiased.

Tufto *et al.* (1996), however, did not examine the behavior of the approximation based on (A1) as between-population differentiation increases in any great detail. We have now found the solution of (18) in Tufto *et al.* (1996), in situations with large between-population differentiation, to sometimes result in matrices $\mathbf{C}$ having negative determinants; *i.e.*, matrices obtained using the proposed method are not always proper covariance matrices. Even though the method was originally intended to work only in situations with limited differentiation, this deficiency of the method is clearly unsatisfactory, as it makes the log likelihood undefined in some parts of the parameter space. We have therefore developed a third, somewhat less complicated, approximation.

As noted in Tufto *et al.* (1996), it is the dynamics of the process in the neighborhood of the observed mean gene frequency that is important in determining the conditional covariances. The third approximation is constructed by assuming that the stochastic elements of the column vector $\mathbf{e}$ in (1) have constant variances, equal to $\bar{p}(1 - \bar{p})/2N_{\mathrm{i}}$. This leads to the same matrix equation (2) as for the original process, except that the elements of $\mathbf{E}$ are replaced by

$$e_{ij} = \begin{cases} \dfrac{1}{2N_i} & \text{for } i = j \\ 0 & \text{for } i \neq j. \end{cases} \qquad \text{(A2)}$$

If we also assume that the elements of $\mathbf{e}_t$ are distributed normally, we have a multivariate autoregressive process that is known to have the multivariate normal as its stationary distribution. The unconditional and conditional covariance matrix of the vector $\mathbf{y} = \mathbf{K}\mathbf{p}$, where

$y_i = p_i - \bar{p}$, $i = 1,\ldots,n - 1$, is then given by the simple transformation

$$\mathbf{C}_{y|\bar{p}} = \mathbf{KCK}. \tag{A3}$$

By computing the conditional likelihood based on an approximation that replaces the original process with a process having constant genetic drift, we hope to obtain reasonably unbiased estimates, at least in situations with a limited amount of genetic differentiation. The performance of the method in any other particular situation will have to be checked by bootstrapping techniques.

It might be mentioned that matrix equation (2) using (A2) can be solved straightforwardly in S-plus after diagonalization of **M** and $\mathbf{M}^T$ and some rearrangement, with operations involving only ($n \times n$) matrices, not suffering from the rapid increase with n in computing time required to solve (2) and (18) in Tufto *et al.* (1996).