Allele-Sharing Methods for Estimation of Population Size

Hans Julius Skaug
Institute of Marine Research, Box 1870 Nordnes, 5817 Bergen, Norway
email: skaug@imar.no

Summary. Genetic data are becoming increasingly important in ecology and conservation biology. This article presents a novel method for estimating population size from DNA profiles obtained from a random sample of individuals. The underlying idea is that the degree of biological relationship between individuals in the sample reflects the size of the population and that DNA profiles provide information about relatedness. A pseudolikelihood approach is taken, involving pairwise comparison of individuals. The main field of applications is seen to be catch data, and as an example, the method is applied to DNA profiles (10 microsatellite loci) from 334 North Atlantic minke whales. It is concluded that the sample size is too small for the method to give useful results. The question about the required sample size is investigated by simulation.

Key words: DNA profiles; Minke whales; Pseudolikelihood.

1. Introduction

The problem of estimating population size has a long history in statistical ecology. Commonly used techniques are mark-recapture methods (Seber, 1982) and distance sampling methods (Buckland et al., 1993). Recently, estimation of population size from genetic data has been considered (Schwartz, Tallmon, and Luikart, 1998). Such data are becoming increasingly available as progress is made in the field of molecular genetics. The methods reviewed in Schwartz et al. (1998) are based on the idea that genetic drift causes linkage disequilibrium and deviation from Hardy–Weinberg equilibrium. Since the effect of genetic drift increases as the population size decreases, these methods are best suited for smaller populations. Being based on a different idea (that of allele sharing among individuals), the method of the present article does not have this limitation. The method is outlined using an analogy to the traditional mark–recapture method.

The Petersen mark–recapture method for estimating population size \( N \) is defined as follows. A random sample of size \( n_1 \) is taken from the population. The animals in the sample are tagged for future identification and then returned to the population. At a later occasion, a second sample of \( n_2 \) animals is taken, and it is found that \( m \) of these have a tag. The Petersen estimator is \( \hat{N}_P = n_1 n_2 / m \). Obviously, this approach does not work if the sampling method is lethal. This problem is relevant when analyzing catch data from commercially exploited populations or data on bycatch from populations that are harvested indirectly.

The setting of the present article is that a single sample of size \( n \) is taken from the population. If it was possible to identify biological relationships between individuals, one could start thinking of a one-sample mark–recapture method. A recapture of an individual would then mean the presence of a close relative in the sample. Genetic data can provide information about biological relationships between individuals. A familiar example is DNA fingerprints (subsequently called DNA profiles) used to identify putative fathers in parent dispute cases. The basic idea is that, on average, an individual shares more alleles with a parent than with a biologically unrelated individual. The same idea can be extended to other types of relationships. It is worth pointing out why the suggested approach would necessarily be more complex than the ordinary mark–recapture method:

- Recaptures can be of different types (parent, sibling, cousin, etc.), while for the ordinary mark–recapture method there is only one type of recapture.
- With a realistic number of genetic markers, it is not possible to discriminate perfectly between different types of relationships.
- The one-sample method depends on demographic parameters such as the variance in family sizes.

Almudevar and Field (1999) have developed an algorithm for detecting sibling clusters based on DNA profiles. However, when the information content in the DNA profiles is low (few loci or low heterozygosity), the error rates of such algorithms will be high. The present approach does not involve explicit classification of relationships but rather is based on the (pseudo) likelihood function of the observed DNA profiles. It may be useful to initially list the assumptions that underlie the rest of the article:

(1) random sampling (individuals are sampled independently with the same probability),
(2) Hardy–Weinberg equilibrium and linkage equilibrium (defined in the next section),
(3) random mating.

A discussion of the plausibility of these assumptions for natural populations is given in Section 6, but we already at
this stage comment on the apparent contradiction between Assumption 2 and the fact that \( N \) is finite. Strictly speaking, Hardy–Weinberg equilibrium and linkage equilibrium are only satisfied in populations of infinite size. However, for large randomly mating populations, Assumption 2 will hold as an approximation. In the same spirit, we will assume that a pair of individuals are either completely unrelated or they are related through a single line of descent. This will not be satisfied in a finite-size population, but since Mendelian segregation acts independently in each generation, DNA profiles from individuals that are not close relatives will be approximately (stochastically) independent. These assertions are verified by simulation in Section 4.

The remainder of the article is organized as follows. Section 2 uses Mendel's law to calculate probabilities for DNA profiles, Section 3 derives the pseudolikelihood function, Section 4 investigates the statistical properties of the pseudolikelihood estimator by simulation, Section 5 applies the method to a dataset consisting of 334 DNA profiles from minke whales, and Sections 6 and 7 provide discussion and concluding remarks.

2. DNA Profiles and Allele Sharing

The only knowledge about genetics needed for developing the present method is the classical chromosome theory, to which Thompson (1988) provides a good introduction. Below, the relevant concepts are reviewed.

Consider \( S \) loci (genes) and denote by \( K_s \) the number of different allelic types existing in the population at the \( s \)th locus, \( s = 1, \ldots, S \). A locus is a particular position on one of the chromosomes. At each locus, an individual holds two alleles, \( (A^{(1)}_s, A^{(2)}_s) \), one maternally inherited and the other paternally inherited. Although there is no information about which allele comes from which parent, we shall treat \( (A^{(1)}_s, A^{(2)}_s) \) as an ordered pair (with the order assigned randomly). For a randomly sampled individual, \( A^{(1)}_s \) and \( A^{(2)}_s \) may be viewed as random variables, each with state space \( 1, \ldots, K_s \). If \( A^{(1)}_s \) and \( A^{(2)}_s \) are independent, the population is said to be in Hardy–Weinberg equilibrium. If, for \( s \neq s' \), the pairs \( (A^{(1)}_s, A^{(2)}_s) \) and \( (A^{(1)}_{s'}, A^{(2)}_{s'}) \) are independent, the loci \( s \) and \( s' \) are said to be in linkage equilibrium. Finally, denote the population frequency of the \( k \)th allele at locus \( s \) by \( x_s(k) \), such that \( x_s(1) + \cdots + x_s(K_s) = 1 \).

Consider a sample of \( n \) individuals. The data on individual \( i \) are

\[
D_i = \left\{ (A^{(1)}_{i,s}, A^{(2)}_{i,s}) \mid 1 \leq s \leq S \right\}.
\]

For natural populations, the joint distribution of \( D_1, \ldots, D_n \) is likely to be complicated, and hence a full likelihood approach for estimation of \( N \) does not seem feasible. Under the independence assumptions made, the marginal distribution of \( D_i \) does not depend on \( N \), so the dependency between the \( D_i \)'s cannot be ignored completely. As a compromise, a pseudolikelihood approach is taken, involving only the distribution of pairs \( (D_i, D_j) \) of DNA profiles. The methods reviewed by Schwartz et al. (1998), on the other hand, are based only on the marginal distribution of the \( D_i \).

2.1 Kinship Coefficients

The distribution of \( (D_i, D_j) \) depends on the relationship between individual \( i \) and individual \( j \). It is convenient to introduce the so-called kinship coefficients, \( \Phi_{ij} \), used in pedigree analysis as a measure of relatedness between two individuals (Thompson, 1986, p. 23). Fix an arbitrary locus \( s \) and let \( A_{i,s} \) and \( A_{j,s} \) denote randomly chosen alleles from \( i \)'s and \( j \)'s DNA profiles, respectively. Then consider the event that \( A_{i,s} = A_{j,s} \). When \( i \) and \( j \) are biologically related, this may happen as a result of \( A_{i,s} \) being a physical copy of \( A_{j,s} \) (or vice versa) or because \( A_{i,s} \) and \( A_{j,s} \) are copies of an allele held by a common ancestor of \( i \) and \( j \) (in the relatively recent past). In such situations, \( A_{i,s} \) and \( A_{j,s} \) are said to be identical by descent (i.b.d.), which is indicated by the notation \( A_{i,s} \equiv A_{j,s} \). However, when \( i \) and \( j \) are unrelated, the event \( A_{i,s} = A_{j,s} \) may still occur because there are only a finite number of different allelic types in the population. For a given relationship between \( i \) and \( j \), we define \( \Phi_{ij} \) as the probability that \( A_{i,s} \equiv A_{j,s} \). Note that this definition depends only on the type of relationship between \( i \) and \( j \), not on the particular locus chosen or the allele frequencies at this locus. Kinship coefficients for some types of relationships are given in Table 1.

We are working under the assumption that \( i \) and \( j \) are related through maximally one line of descent, with the argument being that the shortest genealogical path between \( i \) and \( j \) will dominate over longer paths. The argument does not hold if \( i \) and \( j \) are full siblings, but the probability of this event is \( O(N^{-2}) \) in a randomly mating population. Hence, the possibility of full siblings can be ignored when \( N \) is large. (Note for comparison that the probability of \( i \) and \( j \) being half siblings is of order \( O(N^{-1}) \)).

2.2 Conditional Bivariate Distributions

From the independence assumptions made (Hardy–Weinberg and linkage equilibrium), it follows that the conditional distribution of \( (D_i, D_j) \), given that \( i \) and \( j \) are unrelated, is

\[
p(D_i, D_j \mid \Phi_{ij} = 0) = \prod_{s=1}^{S} x_s(A^{(1)}_{i,s}) x_s(A^{(2)}_{i,s}) x_s(A^{(1)}_{j,s}) x_s(A^{(2)}_{j,s}).
\]

Under the assumption of a single line of descent, the probability that \( i \) and \( j \) share zero, one, and two alleles (i.b.d.) are \( 1 - 4\Phi_{ij}, 4\Phi_{ij} \), and 0, respectively. For a general kinship

<table>
<thead>
<tr>
<th>Relationship</th>
<th>( \Phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent–offspring</td>
<td>1/4</td>
</tr>
<tr>
<td>Full siblings</td>
<td>1/4</td>
</tr>
<tr>
<td>Half siblings</td>
<td>1/8</td>
</tr>
<tr>
<td>Grandparent–grandchild</td>
<td>1/8</td>
</tr>
<tr>
<td>Unrelated</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1

Kinship coefficients for different types of relationships

The only knowledge about genetics needed for developing the present method is the classical chromosome theory, to which Thompson (1988) provides a good introduction. Below, the relevant concepts are reviewed.

Consider \( S \) loci (genes) and denote by \( K_s \) the number of different allelic types existing in the population at the \( s \)th locus, \( s = 1, \ldots, S \). A locus is a particular position on one of the chromosomes. At each locus, an individual holds two alleles, \( (A^{(1)}_s, A^{(2)}_s) \), one maternally inherited and the other paternally inherited. Although there is no information about which allele comes from which parent, we shall treat \( (A^{(1)}_s, A^{(2)}_s) \) as an ordered pair (with the order assigned randomly). For a randomly sampled individual, \( A^{(1)}_s \) and \( A^{(2)}_s \) may be viewed as random variables, each with state space \( 1, \ldots, K_s \). If \( A^{(1)}_s \) and \( A^{(2)}_s \) are independent, the population is said to be in Hardy–Weinberg equilibrium. If, for \( s \neq s' \), the pairs \( (A^{(1)}_s, A^{(2)}_s) \) and \( (A^{(1)}_{s'}, A^{(2)}_{s'}) \) are independent, the loci \( s \) and \( s' \) are said to be in linkage equilibrium. Finally, denote the population frequency of the \( k \)th allele at locus \( s \) by \( x_s(k) \), such that \( x_s(1) + \cdots + x_s(K_s) = 1 \).

Consider a sample of \( n \) individuals. The data on individual \( i \) are

\[
D_i = \left\{ (A^{(1)}_{i,s}, A^{(2)}_{i,s}) \mid 1 \leq s \leq S \right\}.
\]

For natural populations, the joint distribution of \( D_1, \ldots, D_n \) is likely to be complicated, and hence a full likelihood approach for estimation of \( N \) does not seem feasible. Under the independence assumptions made, the marginal distribution of \( D_i \) does not depend on \( N \), so the dependency between the \( D_i \)'s cannot be ignored completely. As a compromise, a pseudolikelihood approach is taken, involving only the distribution of pairs \( (D_i, D_j) \) of DNA profiles. The methods reviewed by Schwartz et al. (1998), on the other hand, are based only on the marginal distribution of the \( D_i \).
coefficient, \(\phi\), we thus have
\[
p(D_i, D_j \mid \Phi_{ij} = \phi) = p(D_i, D_j \mid \Phi_{ij} = 0) \prod_{s=1}^{S} \left[ 1 + \phi \{U(i,j,s) - 4\} \right],
\]
where
\[
U(i,j,s) = x_s \left( A_{i,s}^{(1)} \right)^{-1} \{ I(A_{i,s}^{(1)}, A_{j,s}^{(1)}) + I(A_{i,s}^{(1)}, A_{j,s}^{(2)}) \} + x_s \left( A_{i,s}^{(2)} \right)^{-1} \{ I(A_{i,s}^{(2)}, A_{j,s}^{(1)}) + I(A_{i,s}^{(2)}, A_{j,s}^{(2)}) \}.
\]
(2)

Here, \(I(\cdot, \cdot)\) is the indicator function defined by \(I(u,v) = 0\) if \(u = v\) and is zero otherwise. A derivation of (1) is given in Appendix A.1.

The likelihood ratio statistic for testing the hypothesis that \(i\) and \(j\) are unrelated against the alternative \(\Phi_{ij} = \phi\) is
\[
L_{ij}(\phi) = \frac{P(D_i, D_j \mid \Phi_{ij} = \phi)}{P(D_i, D_j \mid \Phi_{ij} = 0)} = \prod_{s=1}^{S} \left[ 1 + \phi \{U(i,j,s) - 4\} \right].
\]
(3)

The logarithm of \(L_{ij}(\phi)\), known as the LOD score, has been used by Prodohl et al. (1998) to detect parent-offspring relationships (\(\phi = 1/4\)) in an armadillo population and by Herbinger et al. (1997) to detect half and full sibships between cod larvae.

3. Estimation of Population Size

The components of the mechanism generating \(D_1, \ldots, D_n\) are

(1) the population dynamic process generating the genealogy (family tree) of the individuals in the population,

(2) the Mendelian mechanism of gene heritage,

(3) the sampling scheme used to sample individuals from the population.

In the first stage, the genealogy of the population evolves through time as a stochastic process. The population size \(N\) is an important parameter of this process. Other demographic parameters, such as the variance in the (per female) life-time number of offspring, will have status as nuisance parameters. It is assumed that external data are available for estimation of nuisance parameters, and the present article deals only with estimation of \(N\). In the second stage, a gene flow is superimposed on the realized genealogy according to Mendel’s law. For the third stage, random sampling is assumed.

3.1 Unconditional Bivariate Distributions

When individuals \(i\) and \(j\) are sampled from the population, the kinship coefficient \(\Phi_{ij}\) is a random variable. Under the assumption that \(i\) and \(j\) are either completely unrelated or they are related through a single line of descent, the state space of \(\Phi_{ij}\) is \(R = \{2^{-m}, m \geq 2\} \cup \{0\}\). For a given population structure, let \(q_N(\phi)\) denote the probability that \(\Phi_{ij} = \phi\). The subscript \(N\) indicates that the probability depends on the population size. The unconditional distribution of \((D_i, D_j)\) is
\[
p_N(D_i, D_j) = \sum_{\phi \in R} p(D_i, D_j \mid \Phi_{ij} = \phi)q_N(\phi),
\]
where \(p(D_i, D_j \mid \Phi_{ij} = \phi)\) is given by (1). Typically, only a few terms need to be included in (4) because \(p(D_i, D_j \mid \Phi_{ij} = \phi) \approx p(D_i, D_j \mid \Phi_{ij} = 0)\) when \(\phi\) is small. In other words, a DNA profile based on a realistic number of loci does not contain much information about distant relationships. To retain \(q_N(\cdot)\) as a proper probability distribution, the probability corresponding to truncated terms is added to the unrelated category (\(\Phi_{ij} = 0\)).

3.2 Pseudolikelihood

The pseudo-log-likelihood function for estimating \(N\) is
\[
I(N) = \sum_{i < j} \log \left( p_N(D_i, D_j) \right).
\]
(5)

The advantage of a pseudo-log-likelihood approach over a full likelihood approach is simplicity because (5) only involves the distribution of pairs of \(D\)’s. The disadvantage is the loss of information due to ignorance of higher order dependency structures. However, when \(n\) is small compared with \(N\), the information loss will not be large.

The estimator \(\hat{N}\) is obtained by maximizing \(I(N)\), which in practice must be done iteratively. Because \(I\) is not a proper log-likelihood function, an approximation to the variance of \(\hat{N}\) cannot be obtained directly from the second derivative of \(I(N)\), as in ordinary likelihood inference.

The expression for \(I\) depends (through equation (2)) on the population allele frequencies. In practice, these must be replaced by the sample allele frequencies. Throughout the rest of the article, unless otherwise stated, we denote by \(l\) the (log) pseudolikelihood constructed from estimated allele frequencies.

The term \(\sum_{i < j} \log(p(D_i, D_j \mid \Phi_{ij} = 0))\) does not depend on \(N\) and can be subtracted from \(I(N)\) without changing \(\hat{N}\). This trick simplifies the notation because \(I(N)\) can then be expressed compactly in terms of the likelihood ratios \(L_{ij}\).

Below, two special cases of (5) are examined.

Parent–offspring relationships. For a randomly sampled individual, let \(\mu_m\) and \(\mu_f\) denote the probabilities that its mother and father are alive, respectively. It follows from symmetry considerations that \(q_N(1/4) = 2N^{-1}(\mu_m + \mu_f) + O(N^{-2})\). When ignoring all other types of relationships, i.e., assuming \(q_N(1/4) = q_N(0) = 1\), it follows that (5) becomes
\[
l_1(N) = \sum_{i < j} \log \left[ 1 + N^{-1}2(\mu_m + \mu_f) \{L_{ij}(1/4) - 1\} \right],
\]
(6)

where terms of order \(O(N^{-2})\) have been left out. Denote by \(\hat{N}_1\) the maximizer of \(l_1(\cdot)\). To get a feeling for what this estimator does, we can study the behavior of \(\hat{N}_1\) as the amount of genetic information per individual goes to infinity. (Strictly speaking, the following argument assumes that true allele frequencies are used in the construction of \(l_1\).) It is shown in Appendix A.2 that
\[
\hat{N}_1 \to m^{-1}(\mu_m + \mu_f)n(n-1) \quad \text{as} \quad S \to \infty,
\]
(7)
where \(m\) denotes the (unobserved) number of true parent–offspring pairs in the sample and the convergence is in probability. This stochastic limit has the interpretation of being the moment estimator obtained by equating \(m\) to its expected value \(N^{-1}(\mu_m + \mu_f)n(n-1)\).
Population Size Estimation Using Allele Sharing

Table 2

<table>
<thead>
<tr>
<th>Allele</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>37</td>
<td>11</td>
<td>2</td>
<td>17</td>
<td>2</td>
<td>32</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>212</td>
<td>39</td>
<td>4</td>
<td>67</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>130</td>
<td>1</td>
<td>4</td>
<td>22</td>
<td>48</td>
<td>23</td>
<td>3</td>
<td>19</td>
<td>117</td>
</tr>
<tr>
<td>4</td>
<td>226</td>
<td>1</td>
<td>85</td>
<td>52</td>
<td>228</td>
<td>180</td>
<td>31</td>
<td>92</td>
<td>93</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>221</td>
<td>8</td>
<td>72</td>
<td>122</td>
<td>281</td>
<td>142</td>
<td>1</td>
<td>217</td>
<td>268</td>
<td>146</td>
</tr>
<tr>
<td>6</td>
<td>85</td>
<td>2</td>
<td>191</td>
<td>375</td>
<td>62</td>
<td>1</td>
<td>15</td>
<td>101</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>182</td>
<td>132</td>
<td>59</td>
<td>65</td>
<td>24</td>
<td>13</td>
<td>89</td>
<td>56</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>—</td>
<td>10</td>
<td>97</td>
<td>30</td>
<td>1</td>
<td>22</td>
<td>43</td>
<td>184</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>—</td>
<td>26</td>
<td>18</td>
<td>11</td>
<td>—</td>
<td>173</td>
<td>1</td>
<td>29</td>
<td>39</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>—</td>
<td>153</td>
<td>20</td>
<td>1</td>
<td>—</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>—</td>
<td>75</td>
<td>1</td>
<td>1</td>
<td>—</td>
<td>139</td>
<td>1</td>
<td>1</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>—</td>
<td>44</td>
<td>3</td>
<td>—</td>
<td>—</td>
<td>173</td>
<td>1</td>
<td>—</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>—</td>
<td>23</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>51</td>
<td>—</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>—</td>
<td>8</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>5</td>
<td>—</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>—</td>
<td>—</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>—</td>
<td>—</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_s$</td>
<td>.74</td>
<td>.82</td>
<td>.83</td>
<td>.64</td>
<td>.69</td>
<td>.77</td>
<td>.83</td>
<td>.78</td>
<td>.77</td>
<td>.87</td>
</tr>
</tbody>
</table>

**Siblings and grandparents.** Let $\mu_{mm}$ denote the probability that (for a randomly sampled individual) the mother's mother is alive and assume for notational simplicity that all four grandparents have the same probability of being alive. Further, denote by $\mu_{sib}$ the expected number of half-siblings, so that

$$l_2(N) = \sum_{i<j} \log \left( 1 + N^{-1} (2\mu_m + \mu_t) \{ L_{ij}(1/4) - 1 \} + N^{-1} (\mu_{sib} + 8\mu_{mm}) \{ L_{ij}(1/8) - 1 \} \right)$$

(8)

is the pseudolikelihood accounting for relationships with $\Phi \in \{1/4, 1/8\}$. When $S \to \infty$, the maximizer of $l_2(N)$ converges in probability to

$$\frac{2(\mu_m + \mu_t) + \mu_{sib} + 8\mu_{mm}}{m + m'} n(n-1)/2,$$

(9)

where $m'$ denotes the number of pairs in the sample with (true) $\Phi = 1/8$.

**4. Simulation Studies**

The pseudolikelihood function (5) is a $U$-statistic, constructed from dependent random variables (the $D$'s). It is difficult to assess the sampling properties of $N$ analytically, but Monte Carlo simulation can be used. We want to assess

1. the bias and variance of $\hat{N}$ as functions of $N$, $n$, and $S$,
2. the effect of replacing true allele frequencies by estimates in (2),
3. the effect of choosing different truncation points in (4).

To investigate this, data were generated from a modified Fisher–Wright population genetic model:

- Individuals live for exactly two time steps.
- Population size $N$ is constant over time, and hence each generation has $N/2$ individuals (50% males and 50% females).
- An individual picks its mother (father) randomly from the $N/4$ females (males) of the previous generation.

Under this model, $\mu_m = \mu_t = 1/2$, $\mu_{mm} = 0$, and $\mu_{sib} = 4$. For each simulated realization, only three time steps were generated. Alleles were randomly assigned to the first generation according to the allele frequencies in Table 2. The population was sampled in the third time step, yielding individuals from the second and third generation.

The simulation results for $\hat{N}_1$ were as shown in Table 3. The iterative routine used to maximize $l_1(\cdot)$ searched only the interval $[n, 10N]$. Simulation replicas with $\hat{N}_1$ lying outside the interval were discarded when calculating the mean and standard deviation of $\hat{N}_1$. Thus, what is shown in the table are estimates of conditional mean and standard deviation of $\hat{N}_1$, given that $n \leq \hat{N}_1 \leq 10N$. As expected, the standard deviation of $\hat{N}_1$ is a decreasing function of both $n$ and $S$ (for fixed value of $N$).

Denote by $\bar{N}_1$ the version of $\hat{N}_1$ calculated using the true allele frequencies (Table 2) rather than the observed allele frequencies. Table 3 shows the mean and standard deviation of $\bar{N}_1$, calculated from the same simulated datasets as were used for $\hat{N}_1$. For $S = 20$, $\bar{N}_1$ and $\hat{N}_1$ have very similar properties, while for $S = 10$, the sampling distribution of $\bar{N}_1$ is somewhat shifted to the left compared with that of $\bar{N}_1$.

To investigate the performance of $\bar{N}_2$ versus that of $\bar{N}_1$, a second simulation experiment was carried out. For the parameter setting $N = 10,000$, $n = 1000$, and $S = 10$, we generated 1000 datasets. The mean of $\bar{N}_2 - \bar{N}_1$ was 2100, the standard deviation was 500, and the correlation between $\bar{N}_2$ and $\bar{N}_1$ was 0.98. In this experiment (and in the construction of Ta-
Table 3
Mean (E) and standard deviation (SD) of the estimators \( N_1 \) and \( \bar{N}_1 \) (see text) approximated by simulation. The subscripts that appear in the table show the number of simulation replicas for which the estimator fell outside the interval \([n, 10N]\). The number of simulation replicas in the upper part of the table (\( N = 10,000 \)) is 1000, whereas in the lower part of the table, 400 replicas were used.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( S )</th>
<th>( E(\bar{N}_1) )</th>
<th>( SD(\bar{N}_1) )</th>
<th>( E(N_1) )</th>
<th>( SD(N_1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N = 10,000 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>20</td>
<td>11,000( \delta )</td>
<td>11,800</td>
<td>8700( \delta )</td>
<td>10,300</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>10,200( \delta )</td>
<td>5700</td>
<td>8800( \delta )</td>
<td>4400</td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
<td>9700</td>
<td>1600</td>
<td>8900</td>
<td>1400</td>
</tr>
<tr>
<td>( N = 100,000 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>10</td>
<td>101,300</td>
<td>106,300</td>
<td>89,200</td>
<td>88,800</td>
</tr>
<tr>
<td>2000</td>
<td>20</td>
<td>99,800</td>
<td>19,100</td>
<td>99,600</td>
<td>18,900</td>
</tr>
<tr>
<td>3000</td>
<td>10</td>
<td>101,700</td>
<td>45,300</td>
<td>92,100</td>
<td>33,400</td>
</tr>
<tr>
<td>3000</td>
<td>20</td>
<td>101,000</td>
<td>11,000</td>
<td>100,000</td>
<td>11,000</td>
</tr>
<tr>
<td>5000</td>
<td>10</td>
<td>99,000</td>
<td>16,000</td>
<td>93,000</td>
<td>14,000</td>
</tr>
</tbody>
</table>

The major problem is the low sample size, \( n = 334 \). The current estimate of the population size is 112,000 animals (coefficient of variation [CV] 0.10). This is an estimate for the Northeastern Atlantic minke whale abundance.

The second problem is that, although minke whales usually are observed as solitary animals, it is likely that the population on some larger spatial scale exhibits clustering of closely related individuals. Because whales operate within areas of limited size, the assumption about random sampling could be violated. Clustering of related individuals will cause the sample to deviate from Hardy–Weinberg (HW) equilibrium and linkage equilibrium. Thus, equilibrium tests can be used to indicate how severe the clustering problem is. A test of HW equilibrium was performed using the program GENEPOP (Raymond and Rousset, 1995). When applying the Bonferroni correction at significance level 0.05, the hypothesis of HW equilibrium was rejected only for one locus (\( p \)-value 0.0013). Since the method of the present article assumes HW equilibrium, one can argue that this locus should be discarded, but since the \( p \)-value is close to the (Bonferroni corrected) significance level, we decided to include all 10 loci in the analysis. Similarly, GENEPOP was used to test pairwise linkage equilibrium for all the 45 different combinations of loci. Only for one locus pair was the hypothesis of linkage equilibrium rejected when the Bonferroni correction was applied.

Minke whales (males and females) are mature at about 7 years (Olsen, 1997). From a sample of 60 minke whales caught in 1999 (believed also to be representative for the 1996 catches), a mean age of 12.5 years was estimated. This indicates that, on average, no more than two generations are coexisting in the population. Thus, we have used \( \mu_m = \mu_f = 1/2 \), \( \rho_{mm} = 0 \) in the analyses, which correspond to the Fisher–Wright model of the previous section. It is only the value of the \( \mu \)'s that enter into the pseudolikelihood, so to define the estimator \( N \), the remaining assumptions of the Fisher–Wright model (discrete generations, e.g.) are not needed.

The maximizers of \( L_1(N) \) and \( L_2(N) \) were \( \hat{N}_1 = 53,004 \) and \( \hat{N}_2 = 142,012 \), respectively. Parametric bootstrap, with data generated from the modified Fisher–Wright model with \( N = N \), was used to generate 1000 bootstrap replicas \( \hat{N}^{(1)}, \ldots, \hat{N}^{(1000)} \). When constructing the pseudolikelihood from a bootstrap sample, the allele frequencies \( x \) were estimated from the bootstrap sample itself, not from the original sample. The approximate 95% bootstrap confidence interval \([2\hat{N} - u_0, 2\hat{N} - u_0]^{(0.025)}\) was used, where \( u_0 \) denotes the \( \alpha \)-quantile in the empirical distribution of \( \hat{N}^{(1)}, \ldots, \hat{N}^{(1000)} \). The resulting interval based on \( \hat{N}_1 \) was \([50,91,000]\), and based on \( \hat{N}_2 \), it was \([252,200]\). Both intervals were left truncated at zero in order to be biologically meaningful.

The fact that both \( \hat{N}_1 \) and \( \hat{N}_2 \) are finite seems to indicate that data contain at least one pair of related individuals. Can this pair be detected? The expected number of parent–offspring pairs in the sample, assuming \( N = 112,000 \) as the true population size, is \( L^{-1/2}(\hat{\mu}_m + \hat{\mu}_f)n(n-1)/2 \approx 1 \). The natural candidate for a parent–offspring pair would be that with the highest value of \( L_{ij}(1/4) \). (From (2) and (3), it follows that \( L_{ij}(1/4) > 0 \) implies that \( i \) and \( j \) share at least one allele per locus, which is necessary for a parent–offspring pair.) In the sample, there are 263 pairs with \( L_{ij}(1/4) > 0 \), and \( \max_{i \neq j} \{ L_{ij}(1/4) \} = 21,280 \). This maximum equals the (upper) 90% quantile in the null distribution of the random variable \( \max_{i \neq j} \{ L_{ij}(1/4) \} \), where by the null distribution it is meant that the sample (from which the \( L_{ij} \)'s are constructed) contains only unrelated individuals. Thus, the evidence for the presence of a parent–offspring pair is not very strong.

6. Discussion
The main objective of this article has been, in addition to present the method, to find the minimum sample size for which the method is practically useful. If a CV of less than 0.15 is taken as the criterion for usefulness, it is seen from Table 3 that the minimum sample size is approximately \( n = 1000 \) for population size \( N = 10,000 \). This holds both for \( S = 10 \) and \( S = 20 \), but increasing \( S \) from 10 to 20 reduces the standard deviation by approximately 40%. When \( N = 100,000 \), the minimum sample size is \( n = 5000 \) for \( S = 10 \) and \( n = 3000 \) for \( S = 20 \). Note that the fraction of the population that must be sampled to obtain a given CV goes down as \( N \) increases.
Another question is for which range of \( N \) the method can be useful. For small \( N \), it may be inappropriate to ignore inbreeding. We showed in Section 4 that, for \( N = 10,000 \), the effect of inbreeding is not large, but for much smaller \( N \), one will have to take inbreeding into account. For very large \( N \), the method should, in theory, work if the sample size is made sufficiently large. The practical limitation then becomes the laboratory costs of processing a large number of DNA samples. Thus, e.g., most marine fish species currently lie outside the scope of the method. Unfortunately, no attempt has been made to determine the required \( n \) as a function of \( N \), outside the interval covered by Table 3.

The main assumptions underlying the method were listed at the end of the Introduction (Assumptions 1-3). For the minke whale data, the most likely violation of Assumption 1 is spatial clustering of biologically related individuals. However, the fact that Hardy–Weinberg equilibrium was rejected only for one locus (out of 10) indicates that clustering is not a very large problem for this dataset. Spatial clustering can occur on many scales, ranging from clusters consisting of a few individuals up to the level where the population consists of two reproductively separated subpopulations. If a model for the cluster structure can be formulated, simulation experiments can be used to assess the resulting bias in the estimator \( \hat{N} \). Intuitively, the effect of clustering will be to bias \( \hat{N} \) downward. In some applications, such as animal conservation, bias in this direction is less critical than a positive bias, and one would be less concerned about the effect of clustering. Another possible violation of the random sampling assumption is that individuals may differ in their probabilities of being sampled. For instance, there could be size selectivity in the catching operations or the catches could be taken during a time period when only a part of the population was present. Assumption 3 about random mating, i.e., individuals are not monogamous, was made to exclude the possibility of full-sibling pairs. When this assumption is known to be violated, the probability of getting a full-sibling pair should be incorporated into \( q_N(1/4) \). Minke whales are usually observed as solitary animals, and on biological grounds, there is no reason to believe that minke whales are monogamous.

An earlier version of the method (Skaug, 1999) was based on an index of relatedness, defined to be the number of alleles shared by individual \( i \) and \( j \), here denoted by \( w_{ij} \) (0 \( \leq w_{ij} \leq 2S \)). Computation of \( w_{ij} \) involves only integer arithmetic, and \( w_{ij} \) is hence faster to calculate than \( L_{ij}(\phi) \). However, simulation experiments showed that the approach is substantially less statistically efficient than the method of the present article. The reason is that \( w_{ij} \) only depends on \( D_i \) and \( D_j \) and not on the allele frequencies (\( x \)) that go into the calculation of \( L_{ij}(\phi) \). Loosely speaking, the approach based on \( w_{ij} \) is less powerful because it ignores the fact that the sharing of a rare allele provides stronger evidence for relatedness than the sharing of a common allele.

If DNA profiles are collected using a nonlethal sampling method, traditional recaptures (pairs of identical DNA profiles) may occur. It is possible to modify the method to account for the possibility of sampling the same individual more than once. This can be done by extending the set of relationships \( R \) with the value \( \Phi = 1/2 \), which is the kinship coefficient of the identity relationship. The apparatus of Section 3 can still be used. If applied to a dataset resulting from multiple distinct stages of sampling, the method would be an extension of the ordinary mark-recapture method. In situations with few real recaptures, it is likely that this extension would increase the precision of the mark-recapture estimator considerably.

Finally, it should be mentioned that, in a late phase of the preparation of this manuscript, a related article appeared in the literature. Nielsen et al. (2001) also estimate population size from DNA profiles. Their setting is different, however, in that they collect data from mother–calf pairs.

### Conclusion

A method for estimating population size from a single sample of DNA profiles has been presented. We have tried to give guidelines to when the method can be applied and to point out possible pitfalls. Clearly, we still lack a real example where the method has been successfully applied. This article is only a starting point for further developments. The factor limiting the use of the method in practice is the cost of performing genetic analyses. However, these costs are going down rapidly, giving some faith that the method will be a useful tool in the future. In addition to DNA profiles, the method requires knowledge about the demographic structure of the population. This may be the real barrier to the application of the method.

### Acknowledgements

I am grateful to Tore Schweder, who originally suggested that the minke whale DNA register could be used to estimate population size and who also pointed out that the approach in Skaug (1999) uses data inefficiently. Thanks are also due to Thore Egeland, Håkon Gjessing, Berit Myhre Duphy, Bjørnar Olaisen, Erik Olsen, Krystal Tolley, and Nils Øien. Also, I would like to thank two associate editors and two referees for their helpful comments. Parts of this article were written while I was visiting the Department of Statistics, the University of Washington, in the second half of 1999.

### Résumé

Les données génétiques deviennent de plus en plus importantes en écologie et biologie de la conservation. Cet article présente une nouvelle méthode d’estimation de tailles de population à partir de profils d’ADN obtenus par échantillonnage aléatoire de quelques individus. L’idée sous-jacente est que le degré de ressemblance entre individus d’un même échantillon reflète la taille de la population, et que les profils d’ADN fournis une information sur ce degré de ressemblance. Une approche par pseudo-vraisemblance basée sur des comparaisons par paires des individus est utilisée. Le principal secteur d’application visé est celui des données de capture, et à titre d’exemple la méthode est appliquée aux profils d’ADN (10 loci avec micro-satellites) de 334 baleines minke de l’Atlantique Nord. La conclusion est que l’échantillonnage est trop petit pour que la méthode puisse donner des résultats utilisables. La question de la taille d’échantillon nécessaire est explorée par simulation.

### References

A.1 Proof

Due to the assumed linkage equilibrium, it suffices to consider a single locus. For notational simplicity, we define $A = (A_1(1), A_1(2), A_2(1), A_2(2))$. Consider first the case that $\Phi_{ij} = 1/4$. When $S = 1$, equation (1) reduces to

$$p(A | \Phi_{ij} = 1/4) = \frac{1}{4} U(i,j)p(A | \Phi_{ij} = 0). \quad (10)$$

To prove (10), note that, when $\Phi_{ij} = 1/4$ (parent–offspring relationship), individual $i$ and individual $j$ share exactly one allele i.b.d. There are then four possible configurations: $A(1)$ i.b.d. $A(1)$, $A(2)$ i.b.d. $A(1)$, or $A(2)$ i.b.d. $A(2)$, each with probability $1/4$. When conditioning on, e.g., $A(1)$ i.b.d. $A(1)$, the probability of the observed data becomes $z(A(1))z(A(2))$, which we write as $z(A(1))^{-1}p(A | \Phi_{ij} = 0)$. Together with the contribution from the three other configurations, this yields (10). In general, the probability that $i$ and $j$ share one allele i.b.d. is $4\Phi_{ij}$. Thus,

$$p(A | \Phi_{ij} = \phi) = 4\phi p(A | \Phi_{ij} = 1/4) + (1 - 4\phi) p(A | \Phi_{ij} = 0),$$

which, when we substitute for $p(A | \Phi_{ij} = 1/4)$ using (10), is seen to be (1) for the case that $S = 10$.

A.2 Asymptotic Properties of $L_{ij}$

Assume that infinitely many polymorphic loci ($K_s \geq 2$) are available. From (6), we have

$$\frac{\partial}{\partial N} l_s(N) = - \sum_{i < j} \frac{N^{-2} 2(\mu_m + \mu_t) \left\{ L_{ij}(1/4) - 1 \right\}}{1 + N^{-1} 2(\mu_m + \mu_t) \left\{ L_{ij}(1/4) - 1 \right\}}.$$

From general theory for likelihood ratios (Billingsley, 1995, p. 471), it follows that

$$L_{ij}(\phi) \rightarrow \begin{cases} \infty & \text{if } \Phi_{ij} = \phi, \\ 0 & \text{if } \Phi_{ij} = 0, \end{cases} \quad S \rightarrow \infty,$$

such that the equation $(\partial/\partial N) l_s(N) = 0$ reduces to

$$\left( \frac{n(n - 1)}{2} - m \right) - \frac{2(\mu_m + \mu_t)}{N - 2(\mu_m + \mu_t)} - m = 0.$$

It is easily verified that the limit (7) is the solution to this equation. By similar arguments, the result (9) can be proven. Since $L_{ij}(1/4) = 4^{-S} \prod_{s=1}^{S} U(i,j,s)$ and $P(U(i,j,s) = 0 | \Phi_{ij} = \phi) > 0$ for $\phi \neq 1/4$, it follows that $L_{ij}(1/4) \rightarrow 0$ as $S \rightarrow \infty$ when $\Phi_{ij} \neq 1/4$. Thus, (7) will in fact hold even if there are other types of relationships (than parent–offspring relationships) present in the data.