Limitations of hypothesis-testing in defining management units for continuously distributed species

KAREN K. MARTIEN*† AND BARBARA L. TAYLOR†

Contact e-mail: Karen.Martien@noaa.gov

ABSTRACT

Estimating the risk to wildlife populations resulting from human-induced mortality relies on adequately defining population structure. For marine populations, including cetaceans, identifying population boundaries is difficult because most species have large continuous distributions with no obvious barriers to dispersal. For many species, the extreme ends of the range differ in morphology, indicating that population structure exists. However, the lack of distributional hiatuses often makes this structure difficult to detect. A common method of defining structure in such situations is to use genetic differentiation as a proxy for limited movement between areas. Genetic analyses of population structure usually take the form of hypothesis testing, which requires the *a priori* definition of hypothesised units and testing for significant genetic differentiation between them. Simulations are used to examine the performance of hypothesis testing to correctly define population structure. Results show that hypothesis testing is likely to lead the researcher to define fewer management units than are necessary to adequately protect local populations from over-exploitation. The need for the development of new methods of defining management units and for rigorous performance testing of all methods applied in a management context is highlighted.

KEYWORDS: GENETICS; CONSERVATION; MANAGEMENT

INTRODUCTION

Many cetaceans are subject to human-induced mortality, either through direct commercial harvest, subsistence harvest by native communities, or incidental mortality due to entanglement in fishing gear. Regulation of human-induced mortality is usually accomplished through the definition of management units, also known as stocks. However, stock definition has proven notoriously difficult (Donovan, 1991), in large part because distributions are large for cetacean species and barriers to dispersal are not obvious. Nevertheless, successful management requires that human-induced mortality limits be based on units that reflect the actual spatial population structure of the species. To illustrate this problem, consider the management of harbour porpoise (*Phocoena phocoena*) within the state of California. Pollutant analyses (Calambokidis and Barlow, 1991) suggest there are two harbour porpoise populations off the California coast, between which dispersal is limited. Most of the human-induced mortality (due to entanglement in commercial fishing nets) is concentrated in the central California population, which is only about half the size (n = 5,732) of the northern population (n = 11,066) (Forney, 1999). Were these two populations managed as a single unit, the number of animals that could be killed would be calculated based on their combined abundance of 16,798. However, since most of those animals would be taken out of the smaller central California population, that population would quickly become depleted and face possible extirpation if dispersal from the northern population is not sufficient to compensate for the excess mortality. To some degree, errors in stock definition can be compensated for by making precautionary adjustments to the data rather than using 'best estimates' (Taylor *et al.*, 2000b). Such precautionary measures are incorporated into the management scheme used to manage harbour porpoises off California. However, even precautionary management schemes are unlikely to succeed in the face of a 200% overestimate of the abundance of the impacted population, as would occur if the northern and central California populations of harbour porpoises were managed as a single unit.

Over the past decade, genetic studies have become a valuable tool in defining units of conservation. The most common method of investigating population structure is to calculate some measure of genetic differentiation between two hypothesised populations and then test to see if the observed differentiation is statistically significant. Many researchers have pointed out several problems with this approach, calling into question its utility in applied studies (e.g. Bossart and Pashley Powell, 1998; Johnson, 1999; Paetkau, 1999; Taylor and Dizon, 1999; Anderson *et al.*, 2000). Nonetheless, hypothesis-testing remains the most common method of using genetic data to investigate population structure. Consequently, it is important to quantify the frequency and magnitude of errors that are likely when hypothesis-testing is used to define management stocks. It is hoped that this quantification will help scientists to better interpret the results of hypothesis tests of population structure and will enable decision makers to better understand the magnitude of bias likely present in such analyses. This paper outlines two of the major difficulties with using hypothesis tests to determine the population structure of marine species. A simulation approach is then used to estimate the probability of defining fewer stocks than there are populations in an area when using a common hypothesis testing method, Analysis of Molecular Variance or AMOVA (Excoffier *et al.*, 1992).

Hypothesis tests of population structure

Defining hypothesised units

Hypothesis tests of population structure require the researcher to construct an *a priori* hypothesis regarding the number and location of population boundaries. If rates of gene flow between populations are low enough to allow the development of a strong phylogeographic signal (i.e. samples from the same geographic area clustering together on a genetic tree), researchers can use gene trees to guide

* *Department of Biology 0116, University of San Diego, La Jolla, CA 92093, USA.*
† *Southwest Fisheries Science Center, 8604 La Jolla Shores Drive, La Jolla, CA 92037, USA.*

boundary placement (e.g. Brown Gladden *et al.*, 1997). However, even demographically trivial levels of gene flow will prevent such a phylogeographic signal from developing (Taylor, 1997; Bérubé *et al.*, 1998). For instance, a minimum spanning network for the harbour porpoise discussed above shows no apparent geographical clustering of samples despite the fact that gene flow between the northern and southern populations is low enough to require separate management of the two populations (Chivers *et al.*, 2002).

In most published studies of population structure, the authors give no justification for the hypothesised units chosen, nor do they specify whether or not alternative hypothesised structures were examined. Thus, a rigorous analysis of the frequency of different strategies for stratifying data is not possible. Nevertheless, based on both examination of the literature and conversations with researchers regarding their methods of defining hypothesised units, it is possible to discern three commonly employed approaches in genetic studies. First, data are often divided on the basis of political boundaries (Graves *et al.*, 1992; Moritz *et al.*, 1997). Second, samples are divided so as to ensure equal sample size among all units. Typically these units are rather large because researchers realise that increasing the number of samples per unit will increase statistical power and therefore make it more likely that they will obtain statistically significant results. A final, and perhaps most common, method of defining hypothesised units is to simply place hypothesised boundaries in areas where there are gaps in the distribution of samples. Since sampling is often difficult, few investigations of population structure have a deliberate sampling design; rather, samples are gathered opportunistically, with the highest sampling effort concentrated in easily accessible areas. This method of dividing samples into units can be particularly misleading when researchers only publish a map of the distribution of samples (which may be patchy and discontinuous) and no map or description of the actual distribution of the species (which may be continuous).

Statistical power of hypothesis tests

The usefulness of hypothesis tests in defining management units is limited by their reliance on finding statistically significant genetic differentiation. The ability to detect genetic differentiation is often hampered by low statistical power, which is the probability of rejecting the null hypothesis of panmixia when it really is false. Statistical power depends in part on the effect size (F_{st}), which in tests of population structure is given by Wright's (1932) formula (modified for mitochondrial DNA [Takahata and Palumbi, 1985] and finite number of populations [Latter, 1973]).

$$F_{st} = \frac{1}{2NdT + 1}$$

where:

N = the effective number of females in the population;
d = the annual dispersal rate; and
T = generation time.

Therefore, statistical power is inversely related to both the abundance of populations and the rate of dispersal between them. Many marine populations, especially those of commercial value, have large abundances, resulting in small effect sizes and limited power to distinguish them through a hypothesis test. In addition, when defining management units we may want to be able to distinguish between populations with dispersal rates as high as a few tenths of a percent per year. While such movement rates are low enough to have relatively little impact on the demographics of a population, they are high enough to prevent much genetic differentiation from developing, again resulting in low statistical power.

The problem of low statistical power has long been recognised and some authors argue that hypothesis-testing approaches in general are not appropriate for applied studies (Johnson, 1999; Anderson *et al.*, 2000). One of the major difficulties in using hypothesis tests to elucidate population structure is the interpretation of non-significant results. While most researchers are well aware that failure to reject the null hypothesis does not mean that the null hypothesis is true, many continue to make the mistake of interpreting a non-significant result from a hypothesis test of population structure as evidence that the region in question 'lacks' structure and should therefore be managed as a single unit. Even when they are correctly interpreted, non-significant results leave the researcher attempting to define management units in an awkward position. Defining units in the face of non-significant results will appear arbitrary, but failing to define management units will result in the entire region being managed as a single unit and is likely to result in under-protection.

METHODS

To emulate the problem of defining management units for marine mammals, a simulation model was used to generate data for which the actual population structure is known. The study focuses on a stepping-stone model where the level of genetic differentiation is controlled by the dispersal rate between adjacent populations. This stepping-stone model results in isolation-by-distance, one of the most common forms of spatial structure in natural populations and should adequately represent the population structure of most coastal marine mammals. Many pelagic species, particularly large, migratory whales, may exhibit more complicated forms of population structure. Thus, estimates of the performance of hypothesis testing may be conservative since population structure may be even more difficult to detect for species with these more complicated structures.

The model used here was developed by Taylor *et al.* (2000a) and is available from the authors upon request. The evolution of mitochondrial haplotypes was tracked in five populations arranged in a linear stepping-stone. The choice to simulate mitochondrial sequence data was made because it is commonly used in studies of population structure and is particularly useful in identifying demographically independent units (Moritz, 1994; Avise, 1995; 2000). However, as discussed below, the results of this analysis should generalise to the use of nuclear markers, such as microsatellites. The populations were allowed to evolve for 200,000 years and the complete haplotype profile (the sequence of each haplotype and its frequency in all five populations) was recorded from the simulation every 500 years for the last 50,000 years, resulting in 100 haplotype profiles for each combination of effective population size (N_e) and dispersal rate (d).

Annual dispersal rates ranging from 0.002 to 0.01 were examined, along with effective population sizes of $N_e = 100$, $N_e = 300$ and $N_e = 1{,}000$ effective adult females. Annual rates of dispersal were focused on rather than the more familiar per-generation gene flow ($N_e m$) because dispersal rate is the critical parameter in determining whether two

populations can be safely managed as a single unit. Taylor (1997) showed for marine mammals that if two populations are managed together but only one is being harvested (as in the case of the harbour porpoise discussed above), dispersal rates in excess of 1 to 3% per year are probably necessary if the harvested population is to escape extirpation. The generation time for the model was four years, so the per generation dispersal rate was four times the annual dispersal rate. Most cases examined for this paper involved dispersal rates greater than one disperser per generation, but were still sufficiently low that if the management objective is to conserve the species' range then the populations should be managed separately. Thus, we chose a difficult test representative of the performance expected when using hypothesis testing to define management units.

For each haplotype profile, 18 samples were chosen at random from each of the five populations, for a total of 90 samples. This represents a typical sample size for studies of population structure. In order to examine the sensitivity of the results to sample size, some of the analyses were repeated with 36 samples from each population, for a total of 180. The samples were divided into two, three or five equally sized units. These represented three different hypothesised structures that a researcher could use when investigating population structure. In one of the structures (five units), the hypothesised boundaries corresponded to actual population boundaries, while in the other two structures (two or three units), the hypothesised boundaries cut through the middle of actual populations. The average pair wise genetic differentiation between adjacent units was calculated for the three hypothesised structures using the statistic Φ_{st}, the analogue of Wright's F_{st} used in Analysis of Molecular Variance (AMOVA; Excoffier *et al.*, 1992). A permutation test (500 permutations) was used to assign a *p*-value to each measure of differentiation and the results used to determine how many management units should be defined. To mimic the decision process researchers are likely to use in defining management units, the greatest number of units for which all adjacent units were significantly differentiated at the 0.05 level were defined. If none of the three hypothesised structures yielded significant results, the entire region was designated a single unit. We feel that this process roughly approximates the approach that most researchers take to the definition of management units for species with cryptic structure. For each trial, the number of units that would be defined under this criterion was determined. This procedure was repeated five times for each of the 100 haplotype profiles for a given combination of dispersal rate and effective population size. The proportion of the 500 trials that resulted in the correct definition of five units was then determined.

In addition to determining the probability of defining the correct number of management units, three other quantities derived from the statistical analyses were also recorded: average *p*-value between adjacent units; average differentiation (Φ_{st}) between adjacent units; and average statistical power to detect differentiation assuming $\alpha = 0.05$. These averages were taken across all 500 trials and across all pair-wise comparisons between adjacent units. For instance, when the samples were divided into five units, there were four pair-wise comparisons between adjacent units. Power was calculated for each of these four comparisons, and the resulting estimates were averaged to obtain an estimate of the average statistical power when the samples were divided into five units. Since the estimates of average *p*-value, average Φ_{st} and power were based on 500 different samples taken from 100 different points in time, they take into account both sampling error and temporal variation in the degree of genetic differentiation (Whitlock, 1992; Taylor *et al.*, 2000a).

RESULTS

The probability that a hypothesis test will result in the definition of the correct number of units for a species with cryptic population structure is quite low (Table 1). For all parameter combinations examined, the correct definition of five units was the least likely outcome (Fig. 1a). When $N_e = 300$ and $d = 0.002$, the lowest dispersal rate examined, power to detect differentiation between a pair of adjacent units was 0.54 when the samples were divided correctly into five units (Fig. 2a). However, in order to define five units, all four pair-wise comparisons had to be statistically significant, which only occurred with a probability of 0.06 (Table 1). As dispersal rate increased, the probability of defining either five or three units declined, while the probability of defining a single unit increased, resulting in a decrease in the average number of units defined (Fig. 1). Results were similar for effective population sizes of 100 and 1,000 (Table 1). Increasing the sample size improved performance, as expected. However, even for the larger sample size examined, the correct definition of five units was the least likely outcome for all but the lowest dispersal rate (Fig. 1b).

Table 1

Probability of correctly defining five management units as a function of the effective abundances (N_e) of the model populations, the annual rate of dispersal between them (d) and the sample size (n).

Annual dispersal rate (d)	N_e				
	100	300		1,000	
	$n = 18$	$n = 18$	$n = 36$	$n = 18$	$n = 36$
0.002	0.27	0.06	0.31	0	0.022
0.004	0.12	0.02	0.066	0	0
0.006	0.07	0.004	0.042	0	0
0.008	0.03	0	0.006	0	0.002
0.01	0.02	0	0	0	0

Statistical power, average *p*-value and average differentiation were also correlated with the number of units into which the samples were divided. The average degree of genetic differentiation between adjacent units, as measured by Φ_{st} (Fig. 3) and power to detect that differentiation (Fig. 2a) were highest when the samples were divided into just two units rather than being correctly divided into five units. Both Φ_{st} and power declined with increasing dispersal, as expected. The average *p*-value showed the opposite pattern: average *p*-value increased with increasing dispersal rate and was consistently lowest when samples were divided into only two hypothetical populations (Fig. 2b).

The relationships between hypothesised structure and genetic differentiation, power and average *p*-value were consistent across all three effective population sizes examined. Both power and the degree of genetic differentiation were highest, and average *p*-values were lowest, when effective population size was low, as expected.

DISCUSSION

Hypothesis testing is likely to result in the definition of fewer management units than there are distinct populations within a region. Only three of the 25 parameter combinations

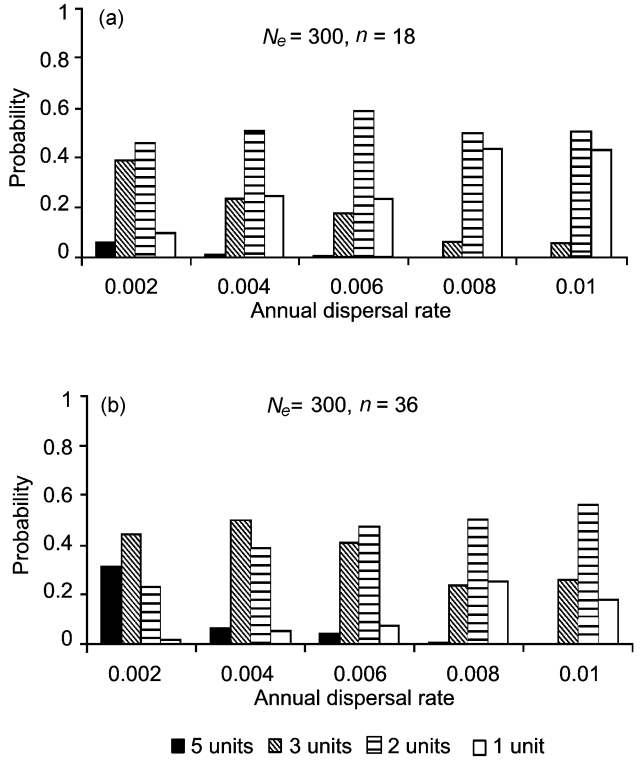

Fig. 1. Probability of defining one, two, three or five units as a function of dispersal rate. The number of units defined was determined by choosing the finest division of the samples that still resulted in significant differentiation between all pairs of adjacent units. Results are for populations with 300 effective adult females with (a) 18 and (b) 36 samples drawn from each population. Results were similar for other effective population sizes (not shown).

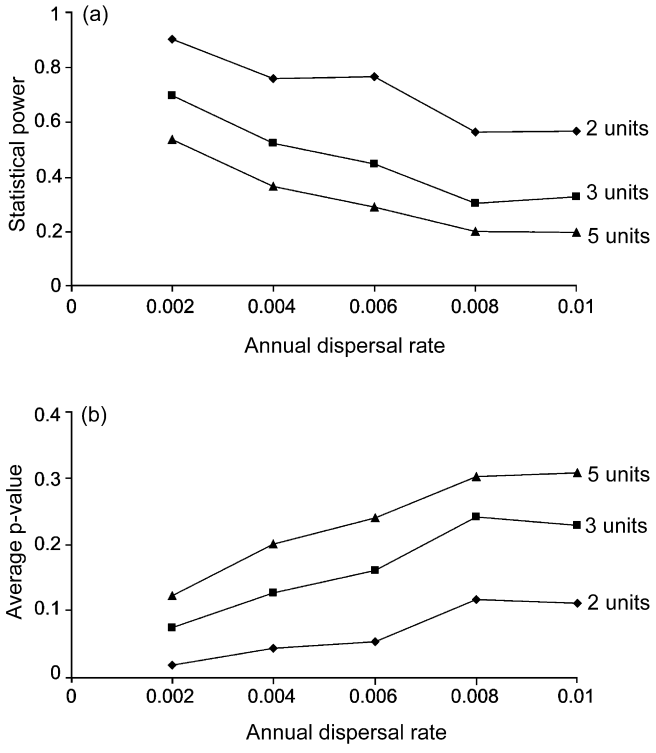

Fig. 2. (a) Power to detect differentiation between adjacent units; and (b) average *p*-value between adjacent units as a function of dispersal rate between adjacent populations when samples are broken into two (◆), three (■) or five (▲) units. Eighteen samples were drawn from each of five model populations arranged in a stepping-stone manner, each with an effective population size of 300 effective adult females.

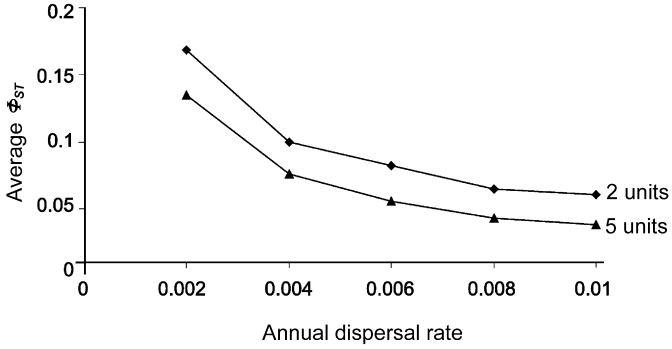

Fig. 3. Average degree of genetic differentiation, as measured by Φ_{st}, observed between units as a function of the dispersal rate between adjacent populations. The model populations had effective abundances of 300 effective adult females. Eighteen samples were drawn from each population and were broken into two (◆) or five (▲) units.

examined resulted in greater than a 10% probability of correctly defining five units (Table 1). Furthermore, the errors made when using hypothesis testing were large, in that for most parameter combinations the researcher would define only one or two units in a region that should be divided into five (Fig. 1). Consequently, hypothesis testing alone is unlikely to result in the definition of management units that adequately protect marine species with cryptic population structure.

The decision criterion used here in deciding how many units to define is probably more conservative than those used by most researchers, in that all four pair-wise comparisons of adjacent units were required to be significant before five units were defined. In reality, many researchers might choose to combine adjacent putative populations that did not show significant differentiation. This approach would not lead to the definition of five units any more often than predicted here, but it would be more likely to lead to the definition of three or four units, reducing the magnitude of the under-protection errors. Nonetheless, based on the results it is hard to imagine any decision criterion that would result in a high probability of defining the correct number of units.

There are two explanations for the inverse relationship between the number of units defined and statistical power. The first, and most obvious, is that by dividing a constant number of samples into more units, you reduce the number of samples per unit. This reduction in sample size reduces the power of the pair-wise comparisons between units (Fig. 2a). However, even when sample sizes are equalised, power is still highest when a region is divided exactly in half, though the disparity is less. Thus, the second explanation is that when a region is divided into only two units, samples from the two extremes of the range are placed in adjacent units. For instance, when the five populations from the computer simulation are divided exactly in half, a comparison is made between a unit containing individuals from population 1 to a unit containing individuals from population 5. Populations 1 and 5 are at opposite ends of the range in question, and thus are maximally differentiated. Comparing units containing these two distant populations inflates the overall degree of genetic differentiation (Fig. 3), resulting in higher power when samples are divided into only two units.

Any series of populations characterised by isolation-by-distance will show this property of increased effect size with decreased number of units. Thus, the results generalise to any model that captures isolation-by-distance, including diffusion models. Indeed, theoretical studies have shown

that stepping-stone and diffusion models produce remarkably similar results (reviewed in Felsenstein, 1976). Similarly, although statistical power was estimated using Φ_{st}, the basic findings that changes in both effect size and per-unit sample size will result in higher power and lower *p*-values when the researcher defines fewer units are robust to the statistic used to measure genetic differentiation.

In species that do not exhibit isolation-by-distance, the inverse relationship observed between the number of units defined and statistical power might not be as strong, because dividing samples coarsely would not necessarily place samples from very distant populations in adjacent units. However, dividing samples coarsely would still increase the number of samples per unit, resulting in increased power. Furthermore, the estimates of statistical power when the samples are divided correctly into five putative populations depend only on the effective abundance of the populations and the rate of dispersal between them, not on the assumption of isolation-by-distance. Consequently, there is no reason to expect better performance when population structure is more complicated than isolation-by-distance. Indeed, the probability of correctly describing population structure through hypothesis testing would probably be even lower than has been estimated here for populations with more complex structure due to the problems associated with correctly stratifying samples *a priori*.

The hypothetical populations simulated for these analyses were all of equal abundance. The problems associated with using hypothesis testing to describe population structure would likely be exacerbated if neighbouring populations differed substantially in size. The populations that need the greatest attention in a management context are those with smaller abundances, since they are the most vulnerable to over-exploitation. While small populations diverge more quickly due to genetic drift, if they are situated next to populations that are substantially larger then the effects of drift can easily be swamped by gene flow from the neighbouring populations, leaving the small, vulnerable populations extremely difficult to detect genetically.

These analyses examined the ability of hypothesis tests to detect differentiation between populations that were in mutation/migration/drift equilibrium. In reality, most natural populations are not in such equilibrium. Rather, populations change through time in concert with their ever-changing environment, resulting in fluctuations in population size, changes in distribution and changes in the rates of exchange with other populations. The impact of non-equilibrium dynamics on our ability to distinguish populations through hypothesis-testing will vary widely. For example, if the abundance of a population has fluctuated through time, the genetic makeup of that population will be much more heavily influenced by its lowest than its highest historic abundance, with the result that it will be much more differentiated from neighbouring populations, and therefore easier to detect using hypothesis-testing, than our analysis would predict. On the other hand, two populations that experience very little gene flow currently but diverged from an ancestral population in the recent evolutionary past will be far less differentiated and far more difficult to distinguish via hypothesis-testing than our results indicate.

While this analysis focused on the use of mitochondrial DNA (mtDNA) data, the conclusions also generalise to the use of microsatellite loci, which are becoming increasingly popular in studies of population structure. Because of higher mutation rates at microsatellite loci and because they are not limited to the use of a single locus, investigations that utilise microsatellite data sometimes have higher statistical power for detecting differentiation than those using mtDNA data. However, in species where dispersal is primarily male-mediated, as is the case for many mammals (Greenwood, 1980), power to detect differentiation will actually be higher for mtDNA due to its strictly maternal inheritance. The effective population size for mtDNA is also four-fold smaller than for nuclear loci, resulting in a larger effect size and higher power to detect differentiation using mtDNA (Avise, 1995). Consequently, in many cases hypothesis testing will be even more likely to result in the definition of too few units when the analysis is based on microsatellite loci rather than mtDNA. Furthermore, the patterns discussed above regarding the number of samples per unit and the average degree of differentiation as a function of the number of units will also apply to microsatellite data. Thus, even when the use of microsatellites does result in an overall increase in statistical power, power and the average degree of differentiation will still be highest when the samples are divided coarsely, into only two units.

Many authors interpret a significant result from a hypothesis test as evidence that the hypothesised structure accurately reflects the underlying spatial structure. The results in this paper show that such an interpretation is not justified. This study has shown that statistical power is highest when a region is divided coarsely, into only two units, even when the boundary defining those units goes through the middle of an actual population. Thus, a finding of significant differentiation across a particular hypothesised boundary does not mean that the hypothesised boundary corresponds to an actual restriction in dispersal. Rather, such a result only indicates that genetic structure is present without lending support for any particular boundary location.

The results of this study highlight two critical needs: the need to both develop better methods to investigate population structure and to subject all methods used in management applications to rigorous performance testing similar to that done here. New methods should move away from the traditional hypothesis-testing paradigm and approach the problem of defining management units from the point of view of parameter estimation and model selection. Given that dispersal rate is the parameter of interest in defining management units, a parameter estimation approach aimed at estimating dispersal rates is likely to be the most fruitful method of defining management units. Critics of hypothesis testing have advocated parameter estimation as a more informative alternative in other applied settings (Johnson, 1999; Anderson *et al.*, 2000). Though it would still require an *a priori* definition of units, such an approach would avoid many of the problems associated with a lack of statistical power that are inherent in hypothesis testing. Pursuing analyses within a parameter estimation framework would provide greater flexibility to managers by allowing them to evaluate the resulting estimates in light of their specific management objectives rather than simply giving them a yes-or-no answer as to whether or not a region is genetically structured, as is the case with hypothesis testing. An estimate of dispersal rate with some measure of uncertainty could also be incorporated quite easily into a formal decision analysis framework. Though analytical approaches are unlikely to result in reliable estimates of dispersal rate (Whitlock and McCauley, 1999), simulation techniques that are free from many of the unrealistic assumptions inherent in analytical methods, such as those of Beerli and Felsenstein (1999; 2001), are likely to be very useful.

Traditional hypothesis-testing approaches to investigating population structure, such as AMOVA, only allow each population structure model to be compared to the null model of panmixia. A model selection approach to defining management units would have the advantage of allowing for direct comparisons between competing models. Some progress has been made in this area. Several new Bayesian and likelihood-based approaches have been published in recent years (Pritchard *et al.*, 2000; Dawson and Belhkir, 2001; Cui *et al.*, 2002). However, these methods have undergone little or no performance testing and none have been tested in a context relevant to management. The results from this study emphasise the need for caution in applying any of these techniques until such performance tests have been completed and have shown that these techniques have a high probability of resulting in the definition of management units that will adequately protect exploited populations. An international programme to develop such a testing framework has recently begun (IWC, 2004).

ACKNOWLEDGMENTS

We would like to thank Susan Chivers, Andrew Dizon, Greg O'Corry-Crowe and Robin Westlake for their contributions to the development of the ideas presented in this paper. Jay Barlow, Ron Burton, Michael Gilpin, Josh Kohn, William Perrin, Trevor Price, Andre Punt and Peter Smouse made helpful comments on earlier drafts of this manuscript. KKM was supported by a National Science Foundation Pre-Doctoral Fellowship and a National Research Council Post-Doctoral Research Associateship while conducting this research.

REFERENCES

Anderson, D.R., Burnham, K.P. and Thompson, W.L. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Manage.* 64:912-23.

Avise, J.C. 1995. Mitochondrial DNA polymorphism and a connection between genetics and demography of relevance to conservation. *Conserv. Biol.* 9(3):686-90.

Avise, J.C. 2000. *Phylogeography: the History and Formation of Species*. Harvard University Press, Cambridge, Massachusetts. 447pppp.

Beerli, P. and Felsenstein, J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152:763-73.

Beerli, P. and Felsenstein, J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl Acad. Sci. USA* 98(8):4563-8.

Bérubé, M., Aguilar, A., Dendanto, D., Larsen, F., Notarbartolo di Sciara, G., Sears, R., Sigurjónsson, J., Urban-R, J. and Palsbøll, P.J. 1998. Population genetic structure of North Atlantic, Mediterranean Sea and Sea of Cortez fin whales, *Balaenoptera physalus* (Linnaeus 1758): analysis of mitochondrial and nuclear loci. *Mol. Ecol.* 7:585-99.

Bossart, J.L. and Pashley Powell, D. 1998. Genetic estimates of population structure and gene flow: limitations, lessons and new directions. *Trends Ecol. Evol.* 13:202-6.

Brown Gladden, J.G., Ferguson, M.M. and Clayton, J.W. 1997. Matriarchal genetic population structure of North American beluga whales *Delphinapterus leucas* (Cetacea: Monodontidae). *Mol. Ecol.* 6:1033-46.

Calambokidis, J. and Barlow, J. 1991. Chlorinated hydrocarbon concentrations and their use in describing population discreteness in harbor porpoises from Washington, Oregon, and California. pp. 101-10. *In:* J.E. Reynolds III and D.K. Odell (eds.) *Marine Mammal Strandings in the United States*. NOAA Tech. Rep. NMFS 98, NOAA NMFS, 7600 Sand Point Way NE, Seattle, WA 98115-0070. 164pp.

Chivers, S.J., Dizon, A.E., Gearin, P.J. and Robertson, K.M. 2002. Small-scale population structure of eastern North Pacific harbour porpoises (*Phocoena phocoena*) indicated by molecular genetic analyses. *J. Cetacean Res. Manage.* 4(2):111-22.

Cui, G., Punt, A.E., Pastene, L.A. and Goto, M. 2002. Bayes and Empirical Bayes approaches to addressing stock structure questions using mtDNA, with an illustrative application to North Pacific minke whales. *J. Cetacean Res. Manage.* 4(2):123-34.

Dawson, K.J. and Belhkir, K. 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.* 78:59-77.

Donovan, G.P. 1991. A review of IWC stock boundaries. *Rep. int. Whal. Commn* (special issue) 13:39-68.

Excoffier, L., Smouse, P.E. and Quattro, J.M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479-91.

Felsenstein, J. 1976. The theoretical population genetics of variable selection and migration. *Annu. Rev. Genet.* 10:253-80.

Forney, K.A. 1999. The abundance of California harbor porpoise estimated from 1993-97 aerial line-transect surveys. Administrative Report LJ-99-02. 16pp. [Available from Southwest Fisheries Science Center, National Marine Fisheries Services, 8604 La Jolla Shores Drive, La Jolla, CA 92037].

Graves, J.E., McDowell, J.R., Beardsley, A.M. and Scoles, D.R. 1992. Stock structure of the bluefish *Pomatomus saltatrix* along the mid-Atlantic coast. *Fish. Bull.* 90:703-10.

Greenwood, P.J. 1980. Mating systems, philopatry and dispersal in birds and mammals. *Anim. Behav.* 28:1140-62.

International Whaling Commission. 2004. Report of the Workshop to Design Simulation-based Perfomance Tests for Evaluating Methods Used to Infer Population Structure from Genetic Data, 21-24 January 2003, La Jolla, USA. *J. Cetacean Res. Manage. (Suppl.)* 6:In press.

Johnson, D.H. 1999. The insignificance of statistical significance testing. *J. Wildl. Manage.* 63(3):763-72.

Latter, B.H.D. 1973. The island model of population differentiation: a general solution. *Genetics* 106:293-308.

Moritz, C. 1994. Applications of mitochondrial DNA analysis in conservation: a critical review. *Mol. Ecol.* 3:401-11.

Moritz, C., Heideman, A., Geffen, E. and McRae, P. 1997. Genetic population structure of the greater bilby *Macrotis lagotis*, a marsupial in decline. *Mol. Ecol.* 6:925-36.

Paetkau, D. 1999. Using genetics to identify intraspecific conservation units: a critique of current methods. *Conserv. Biol.* 13:1507-9.

Pritchard, J.K., Stephens, M. and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945-59.

Takahata, N. and Palumbi, S.R. 1985. Extranuclear differentiation and gene flow in the finite island model. *Genetics* 109:441-57.

Taylor, B.L. 1997. Defining 'population' to meet management objectives for marine mammals. pp. 49-65. *In:* A.E. Dizon, S.J. Chivers and W.F. Perrin (eds.) *Molecular Genetics of Marine Mammals*. The Society for Marine Mammalogy, Lawrence, KS. 388pp.

Taylor, B.L. and Dizon, A.E. 1999. First policy then science: why a management unit based solely on genetic criteria cannot work. *Mol. Ecol.* 8:S11-S16.

Taylor, B.L., Chivers, S.J., Sexton, S. and Dizon, A.E. 2000a. Evaluating dispersal estimates using mtDNA data: comparing analytical and simulation approaches. *Conserv. Biol.* 14:1287-97.

Taylor, B.L., Wade, P.R., DeMaster, D.P. and Barlow, J. 2000b. Incorporating uncertainty into management models for marine mammals. *Conserv. Biol.* 14(5):1243-52.

Whitlock, M.C. 1992. Temporal fluctuations in demographic parameters and the genetic variance among populations. *Evolution* 46:608-15.

Whitlock, M.C. and McCauley, D.E. 1999. Indirect measures of gene flow and migration. *Heredity* 82:117-25.

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. Sixth Int. Congr. Genet.* 1:356-66.

Date received: 16 June 2002.
Date accepted: 3 March 2003.