

Annex I

Report of the Working Group on Stock Definition

Members: Bravington (Chair), Baker, Bickham, Brownell, Jr., Butterworth, Cipriano, Cooke, Cosens, Danielsdottir, Donovan, Givens, Goto, Hammond, Kanda, Kitakado, LeDuc, Lyrholm, Martien, Natoli, Pastene, Paulus, Perrin, Pike, Polacheck, Punt, Schweder, Skaug, Tallmon, Taylor, B., Tiedemann, Walløe, Witting.

1. CONVENOR'S OPENING REMARKS

Bravington welcomed participants, and recalled that an intersessional IWC workshop on the Testing Of Spatial Structure Models (TOSSM) had taken place in La Jolla, USA, 21-24 January 2003. He noted that matters arising from that workshop would likely constitute the main topic of the Working Group's discussions during IWC/SC/55.

2. ELECTION OF CHAIR AND APPOINTMENT OF RAPORTEURS

Bravington was elected as chair. Givens, Kanda, and Perrin acted as rapporteurs.

3. ADOPTION OF AGENDA

The adopted agenda is shown in Appendix 1.

4. REVIEW OF DOCUMENTS

The documents discussed by the WG comprised IWC/55/SD 1,2,3,9,10, SH 8, IST 6 and Rep 3.

5. MODELS FOR POPULATION STRUCTURE AND SPATIAL SIMULATION

5.1 Report from the intersessional workshop (TOSSM)

Chairman's summary

Donovan summarised the report of the workshop to design simulation-based performance tests for evaluating methods used to infer population structure from genetic data (SC/55/Rep3). He explained that the report should still be considered a draft as it is awaiting final comments from participants. The Workshop took place at the Centre for Marine Biodiversity and Conservation, Scripps Oceanographic Institute, La Jolla, California from 21-24 January 2003. Participants from a number of disciplines were present, including population genetics, cetacean biology and management procedures. An important feature was that about one-third of the participants did not have a cetacean background and had not participated in IWC meetings before – this provide an added new dimension to previous discussions on the subject within the Committee. Day 1 therefore comprised background presentations to inform population geneticists of the issue from an IWC perspective and to educate IWC regulars on practical and theoretical issues surrounding population modelling from a genetic perspective. The final 3 days were spent in traditional workshop mode.

From the IWC's perspective, the primary aim was to try and develop ways of examining the performance (particularly in a management context) of existing (and future) genetic techniques that provide stock structure information that feeds into the assessment process. Experience (e.g. in developing *Implementation Simulation Trials* for common minke whales in the North Pacific) has shown that genetic information does not usually provide unequivocal evidence for specific stock boundaries for use in management. Furthermore, few of these techniques have been subject to any form of simulation testing. Even those that have, cannot be considered to have undergone the level of extensive simulating testing to incorporate uncertainty that has been a feature of, for example, the IWC's work on the RMP and AWMP. This is not surprising perhaps, given the scope and complexity of developing suitable genetically-specified simulation datasets.

The purpose of the workshop was thus to begin to develop a suitable simulation framework to allow evaluation of genetic methods used in inferring population structure both in general terms (this has wide implications for many aspects of conservation and management outside the IWC) and from a specifically IWC viewpoint. The specific goals were to:

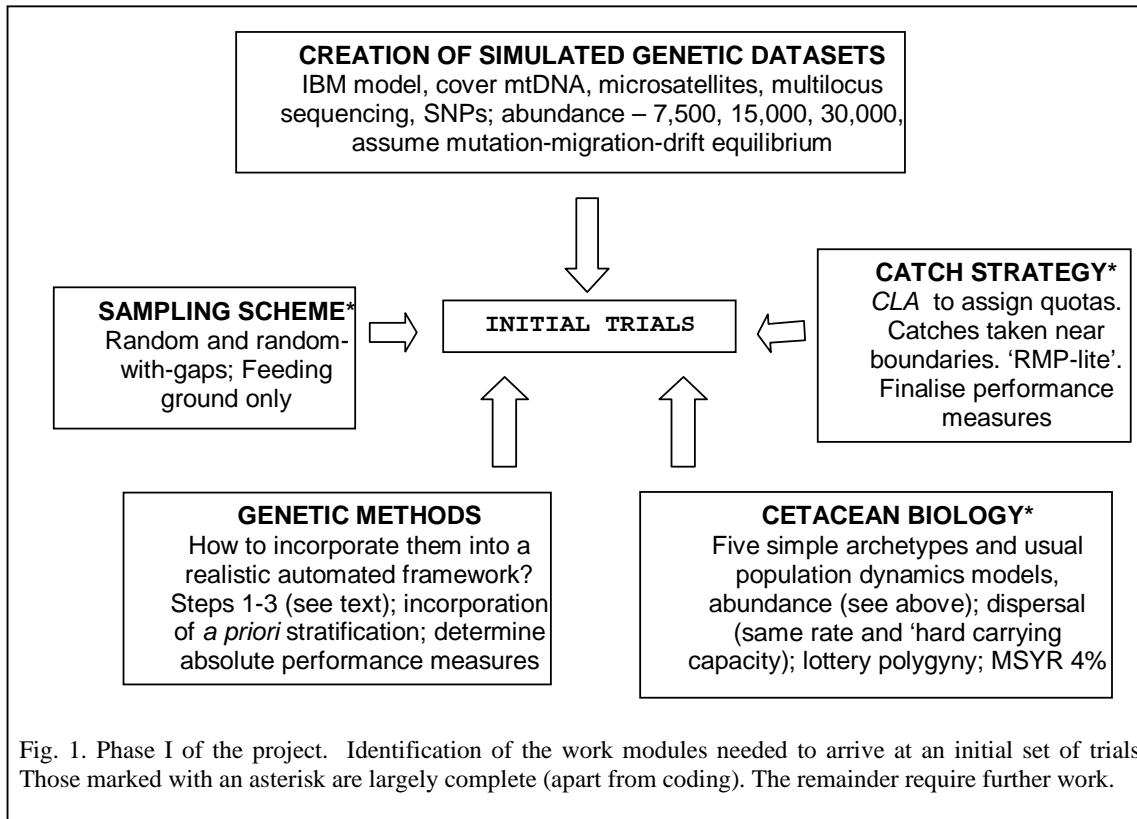
- (1) outline a set of simulations that capture the types of population structure likely to be encountered in migratory baleen whales;
- (2) outline a set of realistic sampling schemes that reflect the types and sizes of samples likely to be available for migratory baleen whale species (e.g., sampling on feeding grounds, breeding grounds, migratory routes, or some combinations);
- (3) design performance measures that reflect the likely performance of the analytical methods with respect to management procedures;
- (4) specify the format of simulated datasets to be used in future work.

It was recognised at an early stage of the workshop that such an ambitious project would inevitably have to proceed in an iterative fashion, as was and is the case in the RMP and AWMP development processes. Given that premise, effort was centred on identifying the conceptual difficulties that needed addressing and developing most detail for Phase I of the process (c.f. *Initial Exploration Trials* in the AWMP process), whilst identifying the types of scenarios that would need to be covered in Phase II and beyond. The modules necessary to make progress are identified in Fig. 1. The figure also includes the key items either decided or left to be decided for Phase I of the project.

Key to the simulation framework is the development and coding /validation of a program to simulate realistic genetic datasets. It was agreed that IBMs (Individually Based Models) represent the most flexible modelling approach but also the most complex and time consuming. Flexibility is particularly important in that it will allow the exploratory work under Phase I to move to Phase II without the need to develop a completely new set of

simulated datasets. Various approaches to developing IBMs in a more efficient way were discussed and an agreed strategy was adopted to find the most efficient method. This is discussed further below.

The question of archetypes was discussed and in Phase I it was agreed to consider five archetypes (see Fig. 2 of SC/55/Rep3). In this exploratory phase it was agreed to choose simple archetypes that would be expected to provide most information on the performance of the methods. In this spirit it was agreed not to include either temporal components or sex-segregation, although these must, of course, be considered in the future. Initially, stock separation will be considered along only one dimension: longitude; latitude; or time. The last will allow for simple consideration of migration. The archetypes are at present confined to baleen whales. This does not imply that consideration of odontocetes is not important but reflects both the need for simplicity (odontocetes have far more complex and variable social systems) and the fact that both the RMP and AWMP apply to baleen whales.



One of the most important difficulties facing the TOSSM project lies in determining an appropriate way to compare the various genetic methods. Not surprisingly, none were developed in order solely to provide direct, automatic recommendations on stock boundaries for the IWC management areas. For example, many require *a priori* stratification based on expert judgement and/or externally chosen 'threshold' values. This is just one way in which the many valuable non-genetic insights into stock structure considerations can be considered (ref.). The issue of how to incorporate 'expert' knowledge in to the simulation framework requires further thought. In fact, the overall issue of how to incorporate quite different methods with different initial objectives requires development. The steps required are relatively easy to define:

- (1) Generate a set of possible boundaries;
- (2) Evaluate these;
- (3) Determine an appropriate set of boundaries for management.

How best to simulate this for methods that were not designed to accomplish each of the steps is an important issue for future work. Possibilities include specific development for each method or the determination of generic approaches for 'missing' steps (specifically (1) and/or (3)). It was agreed to discuss this further in Berlin (see Item 5.3.2)

In order to provide information of the widest possible interest and value, the Workshop agreed to initially investigate the methods at three levels:

- (1) Given a known number of stocks, does a method accurately stratify the data?
- (2) Given a pre-stratified dataset, does a method accurately estimate the number of stocks?
- (3) Given an unstratified dataset and no information of stocks, how well does the method define stocks?

However, (3) is clearly the most likely and most important scenario from a management viewpoint.

Similarly, different performance measures are required to test different issues. Absolute performance measures need to be defined (these may be case-specific and initially specified by 'developers' themselves). These will evaluate how well a method succeeds in achieving what it is designed to achieve. This will be of great general value.

At one level, management performance measures can largely be based on those used in the RMP/AWMP. However, further consideration needs to be given to developing a measure that captures the effect of boundaries on distribution of whaling effort. An important management concept is that of unit-to-serve. This is in practice a policy decision and it was agreed that for simulation purposes, an area will be divided into a sufficient number of cells to allow results to be reported at a variety of scales. Genetic levels of depletion will also be of interest, and it was agreed to consider the number of alleles and the proportion of 'rare' alleles at the start and finish of the management period.

On one sense, the most important aspect of the Workshop was to identify the work modules and work required to enable Phase I to occur. This exploratory phase will improve our understanding of the basic (comparative and absolute) properties of the methods and how they are implemented, as well as providing a feel for the behaviour and interaction of the various factors included in the simulation framework. It will also provide a framework for new methods to be developed (including the combination of various existing methods that at present do not carry out all three of the Steps identified earlier – an area that the workshop recommended be pursued. On the basis of the results of Phase I, Phase II can be developed to provide a more thorough and increasingly realistic examination of methods used to infer stock boundaries in management, perhaps ultimately expanding into more than exclusively genetic techniques.

The major module requiring further work concerns the development of the simulated datasets. The Workshop agreed that this should be coordinated by a small group (Martien (Convenor), Luikart, Tallmon, Taylor and Tiedemann) and recommended Tallman as a likely lead modeller. It was estimated that the work might take 4-6 months and it was agreed that a formal costed proposal should be developed at the Berlin meeting. The other work modules could continue in parallel to this effort and as noted in Fig. 1, many could be carried out within the normal IWC framework including discussion in Berlin. At the same time, an intersessional steering group should be established to ensure that progress is made.

In conclusion the workshop had agreed that an important function of the project was not only to evaluate existing methods but to draw attention to the management and conservation implications of population structure in general and to inspire more development of practical methods.

DISCUSSION

The Working Group thanked Donovan and the Steering Committee of the La Jolla Workshop. The workshop had demonstrated that, although the complexities of setting up the TOSSM genetic simulation framework must of necessity be greater than those in setting up other simulation tests within the IWC, simulation tests are nevertheless feasible in the near future and a clear "roadmap" can be laid out.

The general principals behind TOSSM had been endorsed by the Scientific Committee in 2002. The 2003 Stock Definition Working Group was the first opportunity for scientists not deeply involved in the TOSSM project to comment on the plans developed thus far, and for the project's principals to receive feedback from the broader SC community.

There was considerable discussion of the points raised in Donovan's report. The Working Group confirmed its enthusiasm for this important project, and noted the importance of clearly conveying the motivations to the SC as a whole and to geneticists and cetacean scientists worldwide. In particular, it is important to emphasize that TOSSM is not trying to produce a "black box" procedure to provide automated input to the RMP or AWMP in situations when stock structure is unclear. Rather, it has become evident that it is difficult to devise ways of testing performance of boundary placement measures in "absolute" terms (i.e. purely in terms of biological structure) that still have relevance to management. Absolute measures (e.g. number and location of boundaries selected) can be devised for some very simple population structures, and will be recorded in TOSSM where appropriate, but are difficult to devise or interpret in more complex population structures such as clines. To measure performance for management, it is important to use performance measures of direct relevance to management (e.g. catches, depletion). Such measures can only be evaluated through a simulation framework that incorporates population structure, dynamics, and management, such as TOSSM.

Of course, to implement a management-based simulation, it is necessary to specify not only the boundary-placement method, but also the catch-control rule. It makes sense to use the RMP/CLA for this in TOSSM, both because of its obvious relevance to IWC and because the RMP's behaviour and adjustment mechanisms (catch-cascading etc.) are well-understood. It was noted that the conclusions drawn will inevitably be somewhat specific to the particular catch-control rule chosen. The Working Group **agreed** that, when specifying the details of the simulations, parameters should be chosen deliberately to be informative about comparative performance of different boundary placement methods (e.g. by not bothering to test situations where catches are bound to be negligible whatever the boundaries chosen). In future, it would in any case be straightforward for the IWC or others to adapt the TOSSM programs to use different catch-control rules, for instance in the AWMP or in considering spatial distributions of by-catch.

In summary, the Working Group **strongly endorsed** the TOSSM project as laid out above, and **looked forward** to its development intersessionally.

5.2 Issues requiring further development

The Working Group noted that, of the six modules for phase I of TOSSM identified in section 5.1, three areas would need attention during SC55 and intersessionally. In order of decreasing complexity, these are:

1. Simulating the genetic datasets required as input by boundary placement methods.
2. Adapting methods for studying population structure into boundary placement methods that are suitable for automated simulation testing.
3. Population dynamics, movement and harvesting rules.

5.3 Detailed specifications for simulation testing

5.3.1 Simulating genetic datasets

Paper SD9 proposes a specific framework for TOSSM simulations. TOSSM is designed as a multi-stage project that will start by examining relatively simple population structures and life histories and then progress to more complex scenarios. Therefore, it requires a model that will allow the IWC to quickly start modelling simple scenarios, but retains the flexibility to add complexity at later stages with relative ease. In order to incorporate the level of biological complexity specified at the TOSSM workshop, an individual-based model (IBM) is required. The model must be capable of simulating all four genetic marker types specified in the TOSSM workshop report, and must be computationally efficient if the simulations are to run in a reasonable amount of time.

Many genetic IBMs have already been developed and are available to the public. Using one of these existing models as a starting point for the TOSSM model will greatly reduce development time and cost. After investigating existing simulation models to see which best meet the model

requirements for the TOSSM project, *METASIM* (Strand, 2002) was identified as the best candidate model. This model has several strengths that make it well suited for use in TOSSM. It can already simulate all four types of genetic marker types (mtDNA, microsatellites, SNPS and multiple DNA sequences) identified by the TOSSM workshop as worthy of investigation. *METASIM* incorporates matrix-based demographic projections while retaining the ability to generate individual, multi-locus genotypes, making the model extremely flexible. Finally, the source code for *METASIM*, which is written in C++ and is freely available for use and modification, has a modular structure that allows users to add needed components (e.g., sex-biased dispersal) in a seamless fashion. This is a critical strength of *METASIM* that sets it apart from most other available models.

Before *METASIM* can be used to generate simulated datasets for TOSSM, it will be necessary to validate the model and augment it to incorporate the desired whale-specific features. The first step will be to ensure that, for basic genetic and demographic scenarios *METASIM* provides output that is consistent with theoretical expectations and empirical data. This will ensure that the data generated by *METASIM* are based upon sound genetic and demographic algorithms. The model can currently only simulate unlinked loci and uses a very simple mutation model for nuclear loci; therefore, modifications to the code would be necessary if there is a desire to examine the impact of linkage disequilibrium and alternative mutation models on the performance of different analytical methods.

Another important issue is how *METASIM* will be initialised. The simulations can begin with genotype distributions that are generated by a coalescent model, chosen from observed allele frequency distributions, or randomly assigned to individuals from an infinite allele model. If allowed to run long enough, simulations using all three initialisation methods should eventually come to equilibrium and produce the same results. It is first necessary to check that this happens. Assuming it does, the next step is to determine how long it takes to reach equilibrium values using each of these three initialisations, so that fastest reliable initialisation method can be used for further simulations. Computational time is potentially a serious limiting factor for the large population sizes addressed by TOSSM, so efficiency is very important.

Finally, SD9 noted that before starting to simulate datasets for TOSSM, certain important aspects of whale population biology will need to be incorporated: these include polygyny, sex-biased dispersal, and stage-structured life histories (see SC/55/Rep3). The authors of SD9 plan to check the effects of these added layers of complexity on the time needed to reach demographic and genetic equilibrium values. They will also examine the influence of these demographic parameters on expected population genetic outcomes (e.g., divergence estimates).

The Working Group thanked the authors of SD9 (Tallmon, Martien and Tiedemann) for their efforts, **agreed** that *METASIM* would be an excellent vehicle for pursuing the work of TOSSM, and **strongly recommended** the adaptation of *METASIM* as suggested in SD9.

Concerning timelines, the authors of SD9 expected that the actual programming tasks for *METASIM* in Phase I of TOSSM would take about 4 months. Production of simulated datasets after that will be steady but may be slow, because of the enormous computational demands. Once the datasets are available, actual testing of methods is in principle very fast, subject to the computational demands of the methods themselves.

The Working Group noted that, once *METASIM* has been adapted as proposed, it will be freely available for use by IWC (or anyone) in subsequent generation of simulated datasets. Further modifications might be required in future (i.e. in Phase II of TOSSM), to incorporate additional biological complexities.

5.3.2 Adapting genetic structure methods for use in simulation testing

The process of boundary placement can be conceptualised in three steps, although not all methods will necessarily include sharp distinctions between the steps:

- generate hypotheses about stock structure ("stratification")
- evaluate hypotheses in statistical framework (not necessarily via p-values; could use e.g. estimated dispersal)
- make decisions based on statistical output

Most existing boundary placement methods lack one or more of these steps, and thus are not fully ready for incorporation into TOSSM. The Working Group discussed how to provide standardized versions of steps 1 and 3 that could be applied to any method; this would be particularly useful for methods based on hypothesis tests. Standardized versions would facilitate comparisons of the intrinsic performance of different methods for step 2: of course, though, anybody developing a boundary placement method is free to devise their own versions of steps 1 and 3. Developers might want to consider crosses of different pre-stratification methods with different evaluation methods (step 2). It was noted that not all methods would require pre-stratification; the fundamental data is the spatial position, which will be available to developers and need not be grouped.

When using hypothesis testing to investigate possible boundary placements, it is normal practice to make use of non-genetic information on distribution, morphology, etc. in the pre-stratification. Therefore, an automated pre-stratification will not provide a fair test of how well a hypothesis-testing method would perform in practice. The Working Group discussed the possibility of incorporating the expert judgement aspects of pre-stratification determination, for example by simulating additional non-genetic data. However, it was concluded that this was neither feasible nor necessary for gaining insights into comparative performance, which is the first phase of TOSSM. Results from Phase I will inform how to approach steps 1-3 in Phase II of TOSSM.

After discussion, it was agreed that one reasonable way to proceed would be to start with a moderate number of putative boundaries (~7) evenly spaced across the simulated range, which may or may not be coincident with "real" stock boundaries in the simulation. With this structure, hypothesis tests can be applied to pairwise combinations, as is normal in practice, and boundaries are successively grouped depending on test results. Different options could be considered for deciding when to merge two groups (e.g. threshold p-values, with or without Bonferroni correction; thresholds on estimated dispersal rate). Further suggestions are given in Appendix 2. The Working Group **agreed** that full details could be decided by the TOSSM Intersessional Steering Group recommended by the TOSSM Workshop.

5.3.3 Population dynamics and short-term history; spatial distribution and movement; simulation of harvesting

SC/55/SD3 suggested specifications for extending the standard RMP testing protocol to evaluate alternative genetic methods within an RMP-like management context. These specifications relate to the basic dynamics of the simulated population, the spatial distribution of the stocks (each of which may have site-specificity to a particular set of feeding grounds), how the simulations are initialised and the operating model parameterised, and how future data are generated.

The Group thanked Punt for his contribution, and **agreed** that the framework in SD3 could form the basis for this component of the TOSSM simulation program. Some of the detailed specifications might require reconsideration intersessionally via the TOSSM Steering Group. As noted in section 5.1, it is important to ensure that the management framework being simulated is actually able to provide information on comparative performance of boundary placement methods. Thus, for example, it would be sensible to simulate abundance indices with low CVs, in order to ensure that catches are substantial enough to actually affect population units. Otherwise, the conservatism of the RMP will mean that catches are minimal everywhere, and all methods will perform equally "well".

The Working Group **noted** that the management-related aspects of TOSSM addressed in SD3, which are based partly on existing code would be comparatively easy to incorporate into the METASIM framework, so that METASIM itself effectively becomes the "common control" program for TOSSM simulations.

6. STATISTICAL AND GENETIC ISSUES RELATING TO POPULATION STRUCTURE

Boundary Rank

Boundary Rank (BR) is a hierarchical clustering method designed to generate population structure hypotheses from genetic data. It has been discussed in the past two Scientific Committee meetings. More experience has now been gained with its practical application, and this year the Working Group had a useful and extensive discussion of its behaviour.

Paper SD10 described experiences from several empirical applications of BR. During its development, BR was subject to extensive simulation performance testing which showed that BR performs very well at generating hypotheses that accurately reflect the underlying population structure of the samples being analysed. Nonetheless, when a new analytical method is introduced, it is instructive to also examine its performance when applied to real datasets. BR has now been applied to five datasets in addition to the North Pacific minke whale data: North Atlantic minke whales, South Pacific humpback whales, Western Florida bottlenose dolphins, Eastern Pacific harbour porpoise and Alaskan harbour seals. These analyses are particularly instructive with respect to the performance of BR, since in all of these cases considerable information is already available from non-genetic data and from previous genetic analyses using more traditional analytical approaches. In all five cases, the results of the BR analyses were consistent with previous work, providing empirical validation of the BR method.

Numerous points were raised in discussion, under the following main headings.

Sensitivity

BR requires that samples be grouped *a priori* into initial units, as the method cannot cope with very small sample sizes. Concerns have been raised that results may sometimes be sensitive to the configuration of initial units. It was noted, though, that in the bottlenose dolphin exercise, where naïve and expert analysts had chosen different initial units, the hypotheses ultimately generated by BR were nevertheless similar, conforming to the five bodies of water within the study area.

With respect to optimal sample size for initial units, it was noted that haplotypic diversity is the most important factor in sample size considerations, for any frequency-based genetic analysis. In large and diverse populations, unique haplotypes contribute noise to the analysis. As an indication of this, there is no correlation between grouping order and sample size in BR. It was suggested that using a constant proportion of the population size for initial unit size is a better rule of thumb than using a constant minimum sample size. Historical effects are also relevant, as they may have resulted in greater or lesser haplotypic diversity apart from that determined by population size. B.Taylor commented that the effective sample size can be adjusted by subtracting the total number of haplotypes, to yield a better correlation of statistical power with sample size. In any case, the problem is mitigated by basing decision to aggregate on chi-sq-per-degree-of-freedom, rather than e.g. a p-value.

One suggestion for examining the effect of sample size, is to subsample the larger datasets. Martien agreed this would be useful, noting that subsampling could be incorporated into further work planned with the North Atlantic minke whale data set. Subsampling of simulated datasets had been presented in SC/53/SD7.

BR results are known to be sometimes sensitive to the connectivity matrix, which governs "allowed" patterns of dispersal between pairs of initial units. B.Taylor observed that--- just as with any procedure being used to make inferences from biological data where the answer is not obvious--- effective use of BR requires sensible judgement and biological insight, for example in selecting and trying different connectivity matrices.

Comparison with other genetic analyses, and presentation

Although BR gave generally similar boundaries to *ad hoc* AMOVA-based genetic analyses in SD10, BR also yielded some fresh insights that had not been revealed by AMOVA. For the bottlenose dolphins, several bays had small groups of animals just outside the mouth that would ordinarily have been included in sample strata outside the bays, but which BR grouped with animals in the bays. The BR grouping was consistent with photo-identification data. In Alaska, BR divided harbor seals in the Kodiak Archipelago into two units, a division which had not been contemplated in non-BR analyses but which actually proved to be more congruent with tagging results. It was agreed that that comparisons with non-genetic evidence, such as for two of the five examples in SD10, were of most value in empirical confirmation.

In some cases, of course, different genetic analyses suggest different boundaries. Deciding what to do next is a hard but important issue for management. One problem is that the outputs from most tests of genetic structure do not explicitly show the "estimation uncertainty". In the case of BR, for example, there may be other hierarchies that "fit" the data almost as well; and in the case of hypothesis tests it is very far from clear how to present "uncertainty". For this reason it is hard to tell whether or not different methods are really giving inconsistent answers.

In this regard, it was suggested that some way of quantifying the uncertainty in BR would be particularly desirable, because the BR algorithm is "greedy" and will never reverse an earlier decision to aggregate; therefore it may not reach the best possible hierarchy. It was also noted that the tree diagram used by BR to present the hierarchy can be inadvertently misleading, in that it could be misinterpreted as a phylogeny. Martien commented that alternative graphic aids and optimisation algorithms are being explored, that would address some of these points.

Clines

It is not *a priori* clear how boundaries should be drawn for management of a cline (isolation-by-distance), and a question was raised as to how BR might be expected to perform when given clinal data. It was suggested that pre-analysis of autocorrelations might be used to check for the

presence of a cline. The Working Group noted that this point will be addressed in TOSSM as clines form one of the archetypes to be tested in phase I.

Repeated use of data

Concerns have been expressed in the past about the dual use of data in BR, for hypothesis generation and for hypothesis testing. The Working Group recalled that these concerns have been addressed through the development of permutation tests (SC/54/SD5), at any rate for purposes of hypothesis testing. However, it was noted that there is still a possibility of bias in dispersal rate estimates calculated across boundaries suggested by BR, because BR is likely to place boundaries precisely where estimated dispersal rates happen to be low. Martien said that she had not seen evidence of serious bias in preliminary unpublished simulation tests, but that further investigations were planned.

The North Atlantic Minke Experience: chair's summary

Cooke (chair of NAMWG) related the recent experience of NAMWG in using BR to generate stock structure hypotheses for the RMP Implementation Review for North Atlantic minke whales (Annex D, Appendix 14). BR was applied to the mtDNA dataset derived from the Norwegian DNA register. The total sample size is large (ca. 3500 animals, of which 2600 were used in analysis) but the level of genetic differentiation between subareas is relatively low.

BR picked out the distinction between the Central and Eastern areas, which was also confirmed by direct significance tests. BR also suggested that a subarea in the eastern part of the Central area was separate from both the Eastern area and the remainder of the Central area, but this distinction was not adopted because the subarea in question was thought to be an area of overlap rather than a previously unrecognised stock. BR may have a tendency to identify areas of overlap as being different from both contributing stocks.

BR suggested that the North Sea might be separate from areas to the north. This was confirmed by a test of significance between the North Sea and the area immediately to the north, but areas even further north were found not to be significantly different from the North Sea based on mtDNA from females. Because BR had been run with the North Sea assumed connected only to the area immediately to the north, BR failed to "notice" the similarity between it and areas further away. Nevertheless, the NAMWG had adopted the boundary suggested for the North Sea by BR, because it was supported by circumstantial evidence other than genetics.

BR also suggested a boundary between the Barents Sea and areas to the west. This boundary too was adopted by the NAM Subcommittee, but its exact position was fine-tuned to maximise the genetic distinction across it.

The *p* values for 2-, 3- and 4-stock hypotheses calculated by BR were all non-significant. However, power was expected to be low due to the large number of comparisons made. Consequently the suggested boundaries were retained despite their lack of significance.

Based on this experience, Cooke concluded that BR can be very useful for identifying potential stock structure hypotheses, but that the proposed boundaries do need further close examination case-by-case before being adopted.

Summary

In discussion, the Working Group endorsed Cooke's conclusion, noting that this is consistent with the stated purpose of BR in SC/53/SD7. The Group furthermore welcomed planned developments of BR, including the incorporation of stochastic (non-greedy) optimisation methods, and provision for the use of nuclear markers.

Reproductive autonomy and demographic closure

SH8 attempts to address a central question in the use of standard capture-recapture analysis: what population or population unit is being estimated? Although geographically isolated wintering grounds of humpback whales are assumed to represent breeding units or 'stocks', the evidence that mating occurs in these regions is indirect and the degree of reproductive isolation has not been established. SH8 investigates both the abundance and reproductive autonomy of humpback whales from the New Caledonia wintering grounds using capture-recapture models and paternity inference based on nuclear microsatellite genotyping, mitochondrial DNA sequencing and photographs of natural markings (photo-identification). Microsatellite genotypes were used for an organismal capture recapture estimate of males in the population. The number of inferred paternities among the sampled cow/calf pairs was used as a 'gametic' recapture of males for comparison to the sex-specific organismal population estimate. SH8 assumed that an agreement between these two estimates would be evidence of reproductive autonomy for the New Caledonia wintering grounds, i.e., the proportion of fathers captured by the paternity analysis was consistent with the estimated size of the local male population. The 'gametic' recapture method is similar in approach (although different in intent) to "genetic tagging" or organismal genotype capture recapture estimates (e.g., Palsboll 1999) and estimates of the number of breeding males (Pearse et al. 2001).

The analyses included records of 213 individuals (excluding 16 calves used in paternity inference) identified by genotypes (9 loci), and 210 identified by fluke photographs collected from 1995 to 2001. The 16 cow/calf pairs represent 42% of all cow/calf pairs sited during this seven-year study. A sex-specific estimate of abundance based on genotypes gave comparable numbers of males and females although the variance of the female estimate was greater ($N_m = 382$, $CV = 0.22$; $N_f = 239$, $CV = 0.29$). The likely paternity of 5 calves from 16 cow/calf pairs was assigned to five individual males (one offspring each) from the total sample of 133 non-calf males using the paternity analysis program CERVUS. Confidence in each paternity was high based on probabilities of exclusion and LOD scores calculated by CERVUS. This gametic recapture estimate based on the 5 paternities ($N_m = 379$, $CV = 0.30$) was almost identical to the sex-specific estimate based on the organismal recapture using a two-sample model. The close agreement of the organismal and gametic recapture estimates was considered strong evidence that this small humpback whale wintering ground represents an autonomous population unit that is relatively closed to demographic and reproductive interchange.

The author also drew the Group's attention to the analysis of mtDNA differentiation in SC/55/SH11, which is consistent with SC/55/SH8.

The Group welcomed this novel paper, and recognized its potential utility. An attractive feature is that the approach does not require samples from neighbouring putatively-separate units, and thus could be of particular value where data are limited. It was further noted that the approach allowed permitted reproductive (near-)closure to be established for very small stocks and short timescales, a task which is difficult or impossible for frequency-based methods. Indeed, Baker commented that the timescales may in fact be too short for some management purposes; dispersal events that occur less often than once per generation are liable to be missed.

There was some discussion of the power of the test for closure in SC/55/SH8; it was not clear what effect size (proportion of non-closure) would be ruled out by the data. As well as a p-value, it was suggested that an upper confidence interval on proportion of non-closure might be useful.

It was pointed out that the gametic estimate may be biased if reproductive contributions to the population differ among males (although the five father-calf matches found in SC/55/SH8 did result from five different males), and this bias could be large in such a small sample size; a larger sample size would in principle allow the bias to be corrected, in the same way that heterogeneity of capture probability is handled in mark-recapture analyses. The genetic composition of very closely related neighbouring population could also affect statistical power.

Statistical issues

Paper SC/55/SD1 evaluates the performances of three Bayesian approaches to analysing mtDNA data by means of simulation. The tests involve selecting between one- and two-stock hypotheses when one of these is correct. The "Uninformative Empirical Bayes" method of Cui *et al.* (2002) is more likely to assign high weight to the one-stock hypothesis when it is correct. However, this method also assigns high weight to a one-stock hypothesis when, in reality, a two-stock hypothesis is correct. The performance of a "simple" method based on the posterior distribution for F_{st} is better than that of the "Full Bayes" method of Cui *et al.* (2002) for low sample sizes but this "simple" method does not appear to perform very well for large sample sizes. SC/55/SD1 also examined the sensitivity of the "Full Bayes" estimator of Cui *et al.* (2002) to the choice of the hyper-prior. Although results for log-normal and gamma hyper-priors are consistent in terms of the relative weight assigned to the one-stock hypothesis across data sets involving sub-area 9 of the western North Pacific, there are some differences in the magnitude of the weight assigned, which warrants further investigation.

In the subsequent discussion, the author confirmed that there was no explicit attempt to characterize "effect size" (equivalently, the degree of correlation between allele frequency distributions in adjacent populations that diverged a finite number of generations ago). The task seems computationally quite intractable within the Bayesian framework of SC/55/SD1, at least when large numbers of alleles are involved. Simulation tests could be useful for this, as in TOSSM.

Paper SC/55/IST6 explores the possibility of using Akaike's Information Criterion (AIC) as a means of choosing between stock structure hypotheses. Calculating AIC requires an estimate of the degrees of freedom (DoF) or effective number of parameters required for each model, which can be difficult to quantify for genetic data due to the large number of rare haplotypes often observed. These haplotypes provide little information about population structure, but add parameters to the model, making it less likely that a model containing multiple populations will be chosen by AIC. Three methods of calculating the DoF were examined. In the traditional method, the DoF is the number of populations multiplied by the number of haplotypes less one. However, this assumes that one parameter is required to estimate every haplotype in every population, even though many haplotypes may not be present in some populations. The second approach was to count the number of haplotypes actually present in each population, and then sum across populations. In the third approach, the AIC was calculated by pooling all but the few most frequent haplotypes in the total sample, thereby greatly reducing the number of parameters required for each model.

Performance testing showed that the second approach had the best performance across the range of parameters examined. However, even this approach performed poorly with dispersal rates greater than 20 dispersers per generation, rendering the method of limited use in a management context. Martien hoped that the performance of the method could be improved if a better method of eliminating the 'nuisance' parameters associated with rare haplotypes could be found, and asked the Working Group for input on this problem.

In discussion, members agreed that very rare haplotypes and very common haplotypes are intrinsically less informative about population structure, than haplotypes of intermediate frequency. Several technical suggestions were made for statistical approaches that might alleviate the problem raised in SC/55/IST6. More generally, it was suggested that it is valuable to record not only the unique model with the best AIC score, but also models with scores that are nearly as good. Bravington commented that model selection criteria based on purely statistical grounds, may not be optimal for management.

From a genetic perspective, Tiedemann suggested that use could be made of the genetic relationships between haplotypes, particularly when looking for objective ways to pool rare haplotypes. B. Taylor noted that the typically long-tailed distribution of rare haplotypes can be influenced by a few highly mutable sites, which might be best ignored.

7. UNIT-TO-CONSERVE

In 2002, the Scientific Committee had noted that continued attempts at "stock definition" were not likely to be very useful in the IWC context. This Working Group was instead asked to concentrate on considering possible definitions of "unit-to-serve", and their corresponding implications for management; see JCRM 5 (Suppl.), p49. The Committee had encouraged submission of papers on this issue.

There were no papers on this specific topic available for the 2003 meeting. However, it was noted that paper SC/55/SH8 was highly relevant to one possible unit-to-serve: small groups of animals that are reproductively isolated but perhaps only over short time scales. The Group **acknowledged** the importance of timescale issues in considering unit-to-serve.

The Group further **noted** that the TOSSM project was deliberately structured to allow investigation of how different units-to-serve would respond to different types of management. It was **agreed** that the results from the first phase of TOSSM should help to inform discussions of unit-to-serve in future.

8. WORKPLAN

The Working Group **emphasized** that the TOSSM process, in particular the generation of simulated genetic datasets incorporating at least some biological realism, is essential if progress is to be made in seriously evaluating how well methods for boundary placement are likely to behave in reality, and in further development of new methods. In order to further TOSSM, a substantial amount of intersessional work will be needed on the points in section 5.2 (mainly points 1 and 2). The Working Group **recommended** the formation of an Intersessional Steering Group to co-ordinate these activities (Peter Beerli, Bravington, Donovan, Kanda, Gordon Luikart, Martien, Punt, Tallmon, Skaug, Taylor, Tiedemann) with the following Terms of Reference:

- co-ordinate the work required to implement the TOSSM modules;

- develop an initial set of trial datasets;
- contact developers of genetic methods and provide guidance on linking methods into TOSSM.

Martien agreed to convene the group. A core group (Martien, Tallmon, Tiedemann) will specifically address technical issues arising during development of the METASIM program (SC/55/SD9; section 5.3).

For the 2004 agenda, the Working Group agreed to continue its discussions on TOSSM progress, general statistical and genetic issues related to population structure, and unit-to-serve. Some specific possibilities for the 2004 agenda were also discussed.

Non-genetic data (satellite tags and vocalizations)

In 2001, the Working Group had reviewed various types of non-genetic data in terms of their potential value for studying population structure. This year, several members commented that there has been a substantial recent increase in the quality and quantity of such non-genetic information, principally from satellite tags and also from vocalizations, and that more such data can be expected in the near future. Data on individual movements for other species has proved extremely useful in studying population structure in the context of management. The Working Group **recognized** the importance of this topic, noting also that it may have implications for future directions in TOSSM, and **encouraged** the submission of papers for 2004 showing how such data are, or might be, used in studying and managing populations with possible substructure.

Progress on genetic techniques

The Group's attention was drawn to a recent workshop on the relative utility of SNP and microsatellite data for studying population structure¹. The Working Group **agreed** to discuss the forthcoming report from that meeting in 2004, and welcomed any summaries of case studies.

Time scales in population structure and management

The issue of short-term and long-term demographic patterns has been raised, e.g. in discussion of SC/55/SH8 this year. The Working Group noted that studies of very long-term demographic changes are underway, and that the results will eventually be of great interest to the Committee. It was **agreed** that this would not be a priority topic for 2004, but should be reconsidered for subsequent years when the results of such studies become available.

9. ADOPTION OF REPORT

The report was adopted on Monday 2nd June 2003 at 17:30.

Appendix 1

1. Convenor's opening remarks
2. Election of chair and appointment of rapporteurs
3. Adoption of agenda
4. Review of documents
5. Discussion of models for population structure and spatial simulation
 - 5.1 Review report of intersessional workshop (TOSSM)
 - 5.2 Identify issues requiring further development at this meeting
 - 5.3 Develop detailed specifications for simulation testing
 - 5.3.1 Simulating genetic datasets
 - 5.3.2 Adapting genetic structure methods for use in simulation testing
 - 5.3.3 Population dynamics and short-term history; spatial distribution and movement; simulation of harvesting
6. Statistical and genetic issues relating to population structure
7. Unit-to-serve
8. Workplan
9. Adoption of report

¹ "Technical and analytical methods for wildlife genetics: developments for use of SNPs (Single Nucleotide Polymorphisms)" Workshop in Leipzig, Germany, 11-13 September 2002

Appendix 2

SUGGESTIONS FOR AUTOMATICALLY GENERATING HYPOTHESIZED BOUNDARIES TO SIMULATE SUBJECTIVE *A PRIORI* STRATIFICATION

Karen K. Martien

Background

One of the issues that will have to be addressed in the TOSSM (Testing of Spatial Structure Methods) simulation performance testing project is how to fairly test methods that require *a priori* stratification of samples into hypothesized units. In reality, this step is typically performed subjectively by using whatever biological data may be available to guide the placement of hypothesized boundaries. It was agreed by the Stock Definition Working Group that, at least for the first phase of the project, attempting to incorporate this kind of 'expert judgement' in the performance tests was too complicated. Instead, methods of automatically generating hypothesized boundaries in a way that crudely mimics *a priori* stratification were discussed. The ideas presented in this Appendix are elaborations of suggestions made by members of the SDWG. They are intended to serve as a starting point for the TOSSM steering group in finalizing the specifications of the boundary generating method.

General approach

When deciding how to stratify a dataset, two issues must be addressed: 1) how many hypothesized boundaries to place and 2) where to place them. In both cases, imperfect subjective decision making can be simulated by drawing random numbers from appropriate distributions. The widths of the distributions would represent the amount of auxiliary information available to guide stratification. When substantial biological data are available (e.g., distributional data, habitat differences, morphology, satellite tagging, pollutant data, etc.) stratification would presumably be more accurate than when such data are not available. These two situations would be presented in the boundary generation routine by narrower and wider distributions, respectively.

Number of boundaries

There are several possibilities for choosing the number of boundaries. First, the number of boundaries could be chosen from an unskewed distribution centred around the correct number. This would represent random errors in the number of boundaries. In reality, however, errors are not always random. Some people tend toward defining more hypothesized boundaries ('splitters'), while others tend toward defining fewer ('lumpers'). These behaviours can be modelled by selecting the number of boundaries from distributions that are either skewed or biased in the appropriate direction.

Note that during phase I of the TOSSM project, the 'lumper' strategy would only be relevant to archetypes II and III, as all other archetypes assume only two populations. In those cases, it is not possible to define too few hypothesized boundaries.

Boundary placement

When few data are available to guide the stratification of samples, researchers tend to define relatively few strata of roughly equal size. Therefore, the automatic boundary generation routine should be designed to define roughly equally-sized strata as well.

Most stock structure scenarios being considered in the first phase consist of only two populations. Therefore, in order to meet the requirements that i) the hypothesized boundaries divide the samples into roughly equal units and ii) one of the boundaries should roughly correspond to the correct boundary, the number of strata defined will always have to be a multiple of two.

The procedure for placing boundaries could therefore be as follows:

1. Choose the number of strata by drawing a random number from the appropriate distribution and rounding to the nearest multiple of two.
2. Place the first boundary by choosing a random deviate from a distribution centred on the correct boundary location.
3. For the remaining boundaries, determine the boundary locations that would result in equally-sized strata, and then choose boundary locations from distributions centred on those locations.

This procedure is a very simplistic one appropriate for the first phase of the TOSSM project. More sophisticated approaches may need to be investigated at a later date, especially if archetypes involving more than two populations are considered. The intersessional steering group for TOSSM should also consider an even more simple approach in which the a set number of boundaries placed in each simulation is set at a constant value, probably either five or seven. Such an approach would eliminate the number of boundaries chosen as a source of variation and instead limit the variation to the accuracy of the location of the hypothesized boundaries.

REFERENCES

- Cui G, Punt AE, Pastene LA, Goto M (2002) Bayes and Empirical Bayes approaches to addressing stock structure questions using mtDNA data, with an illustrative application to the North Pacific minke whales. *J. Cet. Res. Manage.* 4:123-34.
- Marshall TC, Slate S, Krusk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology* 7:639-655.
- Palsboll PJ (1999) Genetic tagging: contemporary molecular ecology. *Biol J Linn Soc* 68:3-22.
- Pearse DE, Eckerman CM, Janzen FJ, Avise C (2001) A genetic analogue of 'mark-recapture' methods for estimating population size: an approach based on molecular parentage assessments. *Molecular Ecology* 10:2711-2718.
- Strand, AE (2002) METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics. *Molecular Ecology Notes* 2:373-376.