TOSSM dataset handbook

Introduction

The TOSSM datasets are a collection of simulated genetic datasets developed for use in the Testing of Spatial Structure Methods project. The goal of the TOSSM project is to conduct comparative performance testing of analytical methods for defining management units from genetic data. The purpose of this document is to provide details on how the datasets were parameterized and generated. Further information on the TOSSM project and the TOSSM package are available at http://swfsc.nmfs.noaa.gov/TOSSM.

List of Scenarios

The participants at the TOSSM workshop held in La Jolla, CA, in January of 2003 outlined the parameters to be used in generating datasets for the TOSSM project (IWC 2004). The datasets represent five population structure archetypes:

Archetype I: a single panmictic population
Archetype II: a linear stepping-stone, with dispersal occurring only between adjacent populations. There are four variants of this archetype: a) two populations with equal abundance, b) two populations with unequal abundance (90:10), c) three populations with equal abundance, and d) three populations with unequal abundance (45:45:10).
Archetype III: diffusion-like isolation by distance
Archetype IV: two populations with separate breeding grounds but overlapping feeding ground(s).
Archetype V: A single breeding population with separate feeding grounds. Feeding ground philopatry is learned from the mother.

For each Archetype, there are simulated datasets with carrying capacities (summed across all populations) totaling 2,500, 7,500, and 15,000. Because all simulations are initialized at carrying capacity and are simulated under density dependent population growth, the simulated populations will remain at or very near carrying capacity for the entire simulation.

Annual dispersal rates of 5x10-6, 5x10-5, 5x10-4 and 5x10-3 were chosen to span the range of rates that might be of interest to conservation biologists. The lowest of these rates corresponds to approximately one disperser per generation and will therefore result in populations that are following independent evolutionary trajectories. The highest rate results in populations that are demographically independent, but genetically very similar.

A full list of the scenarios planned for simulation is given in the Appendix. Unlike the other four archetypes, Archetype III assumes continuous genetic variation over space. At the 2003 TOSSM workshop in La Jolla, CA, it was agreed that this archetype could be simulated by creating an Rmetasim landscape that contained many small populations with relatively high dispersal rates between adjacent populations (IWC 2004). The Rmetasim matrices necessary to implement this archetype have not yet been designed, and so no datasets are currently available for this archetype.

Archetype IV can be simulated by simply drawing samples from Archetype II datasets and assigning overlapping spatial coordinates to samples from the two populations. Such sampling can be easily accommodated by the TOSSM package (see the second example for the function run.tossm in the tossm package).

Life history matrices

We used vital rate estimates for gray whales (Table 1; Reilly 1984) to parameterize stage-based matrices for use in generating the TOSSM datasets. We constructed matrices representing vital rates both at carrying capacity and near zero population density. Rmetasim (Strand 2002) implements density dependence by interpolating between these matrices.

Table 1. Vital rates for eastern Pacific gray whales. The first column shows the estimates used by Reilly (1984). The second column includes the pregnancy rate estimate from Perryman et al. (2002) and a juvenile survival rate adjusted to produce zero population growth. The third column shows the assumed biological limits that would be approach near zero population density.

Vital Rate	Estimate used by Reilly	Estimate at K	Value assumed near zero density
Juvenile survival*	0.893	0.92	0.940
Adult female survival	0.946	0.946	0.946
Adult male survival	0.954	0.954	0.954
Age of first reproduction	8 (5-11)	8 (5-11)	5
Pregnancy rate	0.463	0.24	0.5
Λ	1.011	0.997	1.068
Generation Time	19.4	19.4	16.9

**Juvenile survival rate was not empirically estimated, but rather was derived fixing all other vital rates and adjusting to produce the desired value for λ.*

We defined five life-history stages: juvenile1, juvenile2, fertile female, lactating female and adult male (Figure 1). The use of two juvenile stages allowed better control of the average age at first reproduction (AFR). The juvenile1 stage corresponds to animals too young to have reached sexual maturity, while the juvenile2 stage represents animals that are old enough that they could mature and move into an adult stage class in the next time step. The use of separate fertile and lactating stages for adult females allowed enforcement of a minimum 2 year inter-birth interval. We used the above vital rates to parameterize the stage-based matrix using a fixed stage duration model (Caswell 2001). More sophisticated models are available (e.g., variable stage duration, negative binomial stage durations), but require estimates of the variance of the time spent in each stage, which are not available.

The resulting matrices are shown in Table 2. Males and females are undifferentiated prior to sexual maturity. When an individual from the juvenile2 stage class matures, it has an equal probability of moving to the fertile female and adult male stages. The rates of increase (λ), generation times, pregnancy rates and proportion of the population in the lactating female stage-class resulting from these matrices (Table 3) are in close agreement with the values in Table 1.

Table. 2. Stage-based matrices for use at a) zero population density and b) carrying capacity. Stage class abbreviations are juve1 = juvenile1, juve2 = juvenile2, fert = fertile female, lact = lactating female, and male = adult male.

a)	juve1	juve2	fert	lact	male
juve1	0.730	0	0	1.0	0
juve2	0.210	0	0	0	0
fert	0	0.47	0	0.946	0
lact	0	0	0.946	0	0
male	0	0.47	0	0	0.954

b)	juve1	juve2	fert	lact	male
juve1	0.797	0	0	1.0	0
juve2	0.123	0.718	0	0	0
fert	0	0.101	0.648	0.946	0
lact	0	0	0.300	0	0
male	0	0.101	0	0	0.954

Table 3. Vital rates resulting from the matrices in Table 2. Generation time is defined as the average age of the mothers of a cohort. Pregnancy rate is defined as the proportion of adult females that are in the lactating stage class, while % lactating is the proportion of the total population that is in the lactating female stage class.

	Zero Population Density	Carrying Capacity
Age at first reproduction	10	5
Rate of increase (λ)	1.072	0.998
Generation time	17.1	20.0
Pregnancy rate	0.47	0.23
% lactating	0.13	0.06

Landscape initialization

The simulated populations were initialized at carrying capacity and in stable age distribution. The mtDNA haplotype and microsatellite allele frequencies of the simulated populations were initialized from distributions generated by the coalescent model SIMCOAL v2.1.2 (Laval and Excoffier 2004). Initializing from a coalescent rather than from random haplotype and allele frequencies resulted in a considerably shorter burn-in, as the frequency distributions generated by SIMCOAL are already very close to equilibrium.

SIMCOAL requires as input an estimate of effective population size (N_e). It is therefore necessary to estimate N_e for each scenario by running a simulation that contained a single microsatellite locus and a 500 bp mitochondrial DNA sequence, each with no mutation and initialized with 1000 alleles. The simulation was run for 2000 years (100 generations). By calculating heterozygosity at the beginning and end of the simulation, N_e can be estimated using the equation

$$H_t = H_o\left(1-\frac{1}{2N_e}\right)^t$$

where H_t is heterozygosity at time t and H_o is heterozygosity at time zero. We replicated the simulation 10 times and averaged across replicates to obtain a mean N_e for both nuclear and mitochondrial markers. It was necessary to estimate N_e separately for the two marker types because the haploid and uni-parentally inherited nature of mitochondrial DNA results in a markedly smaller effective population size for the mitochondrial genome.

Following initialization, simulated populations were projected forward for 1,000 years (50 generations) in order to ensure that they had reached demographic and genetic equilibrium. Previous examinations of the trajectories of the number of mtDNA haplotypes, microsatellite alleles, and heterozygosity in both markers indicated that 1,000 years was a sufficient amount of time to ensure that these values were relatively stable. A sample of all markers was independently generated from SIMCOAL for each burn-in replicate.

Genetic Marker Characteristics

mtDNA – Mitochondrial DNA haplotypes being simulated in the TOSSM datasets are 500 bp in length and have a mutation rate of 5x10-3 per generation for the full sequence. This mutation rate is based on mutation rate estimates for the mitochondrial control region (Heyer et al. 2001) and produced haplotype distributions consistent with what is seen in many large whale species (Figure 2). For Archetype 1, scenario 1 (one population, N=7,500), the number of haplotypes in the final datasets averaged across replicates was 118 (s.d. = 8.62) (Figure 3).

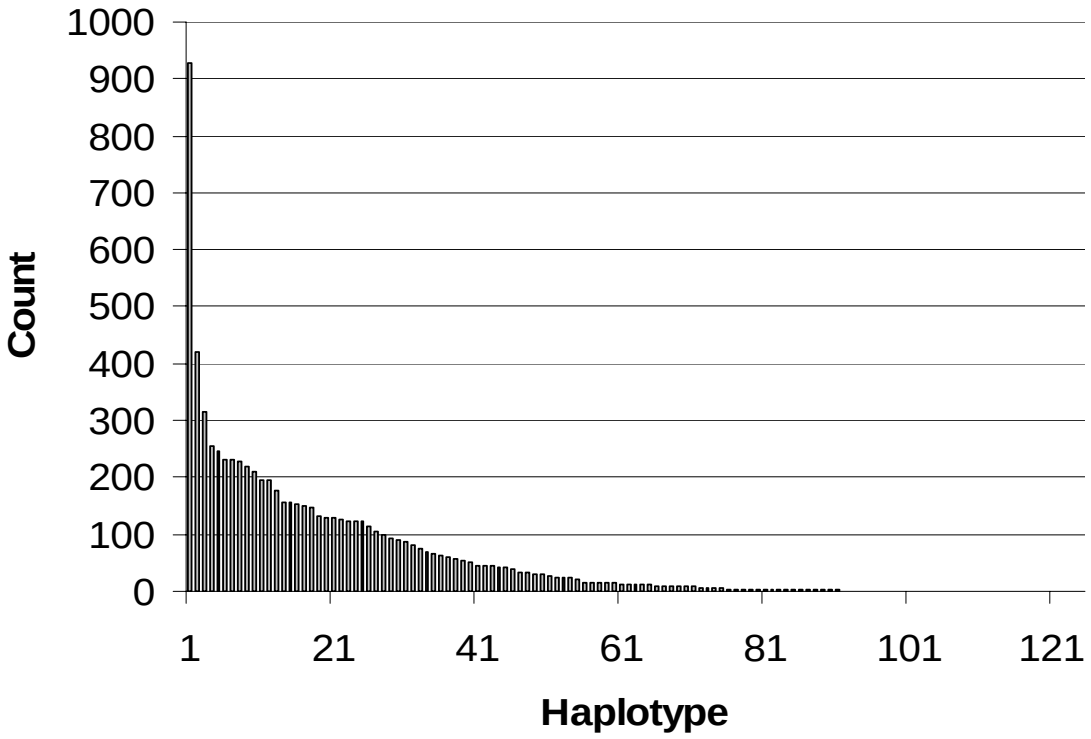

Figure 2. Example haplotype frequency distribution for one replicate of Archetype 1, scenario 1. Haplotypes are order along the x-axis from most to least frequent. The number of individuals with each haplotype (out of 7465) is given on the y-axis. The distribution includes a few very common haploltypes and a long tail of very low frequency haplotypes, as is typically seen in empirical datasets. Note that there were 125 haplotypes in this replicate; haplotypes 92 through 125 were each represented by a single individual and are not visible on the graph.

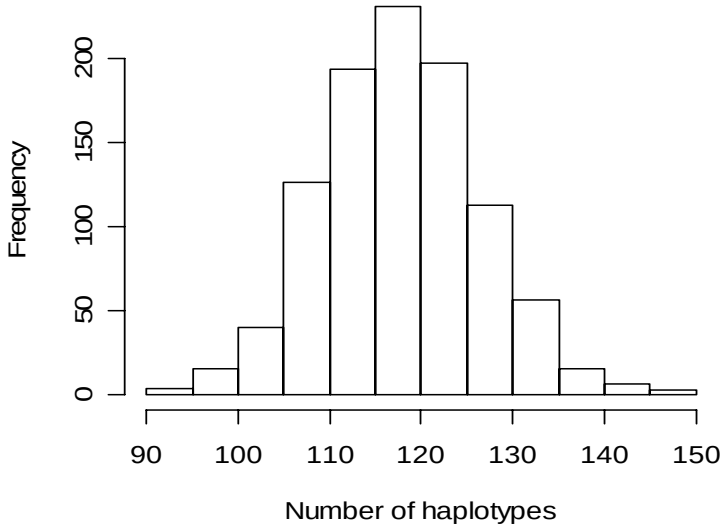

Figure 3. Distribution of number of haplotypes across replicates for Archetype 1, scenario 1.

Microsatellites – The TOSSM datasets use 30 microsatellite loci, each with a mutation rate of 2x10-3. This mutation rate was based on estimates from the literature (see Brohede 2003 for a review) and resulted in allele frequency distributions consistent with those typically seen in studies of population structure (Table 3). The number of alleles per locus ranged from 6 to 26 (Figure 4), which is consistent with the number of alleles typically observed in empirical datasets.

Table 3. Example allele frequencies at all 18 loci for one replicate of Archetype 1, scenario 1.

Allele	Loc1	Loc2	Loc3	Loc4	Loc5	Loc6	Loc7	Loc8	Loc9
1	3276	6349	3798	3920	2813	3873	4237	3785	2487
2	2723	2618	2649	3672	2773	2808	3355	2894	1742
3	2513	2132	2341	1989	2472	1558	3229	2803	1671
4	2008	1881	1982	1919	2397	1344	2144	1680	1584
5	1624	825	1367	1522	2094	1290	1025	1415	1475
6	1400	318	1245	1138	833	1267	685	1169	1308
7	1146	285	725	425	660	1085	244	510	1254
8	237	166	393	277	567	847	11	435	1129
9	3	156	313	68	191	280	0	93	748
10	0	122	106	0	130	259	0	58	714
11	0	61	11	0	0	165	0	49	452
12	0	17	0	0	0	114	0	20	330
13	0	0	0	0	0	40	0	19	28
14	0	0	0	0	0	0	0	0	8
15	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0

Allele	Loc10	Loc11	Loc12	Loc13	Loc14	Loc15	Loc16	Loc17	Loc18
1	5151	3376	3105	2842	3161	4461	6205	3368	4394
2	3715	2652	2961	2096	2054	3503	3048	2468	2749
3	1984	2594	2871	1701	2038	3399	1799	2392	2527
4	1466	1561	1534	1650	1578	1461	1438	1848	1599
5	957	1210	1300	1640	1237	1094	666	1062	1224
6	726	981	1225	1226	1038	383	580	1044	1088
7	517	827	1111	1127	887	346	468	770	614
8	408	469	680	1101	839	268	211	670	579
9	6	459	131	727	734	15	186	561	112
10	0	445	12	464	729	0	168	354	44
11	0	354	0	340	425	0	158	153	0
12	0	2	0	16	72	0	3	99	0
13	0	0	0	0	60	0	0	92	0
14	0	0	0	0	59	0	0	45	0
15	0	0	0	0	10	0	0	3	0
16	0	0	0	0	7	0	0	1	0
17	0	0	0	0	2	0	0	0	0

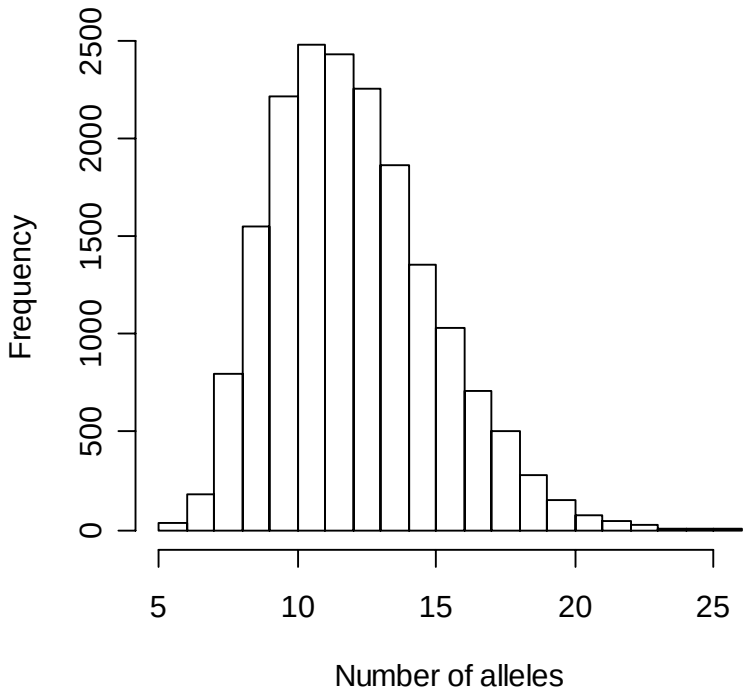

Figure 4. Distribution of number of alleles per locus. Data are combined for all 30 microsatellite loci from all 1000 replicates of Archetype 1, scenario 1.

Appendix: TOSSM Scenarios

Table A1. Specifications for all scenarios specified by the TOSSM steering committee. Archetype 4 is not included in the list, as it can be generated by simply sampling from Archetype 2 simulations. 'Total abundance' refers to the combined abundance of all populations in the dataset. 'Relative abundance' indicates the percentage of the total abundance that is in each of the populations. The last column indicates whether a dataset has been generated as of 1 August, 2007. Scenario numbers are not continuous because some scenarios originally specified have since been eliminated.

Scenario name	Archetype	#populations	Total abundance	Relative abundance	Dispersal	Complete?
Arch1_1	1	1	7500	n/a	0	X
Arch1_2	1	1	15000	n/a	0	X
Arch1_4	1	1	2500	n/a	0	X
Scenario name	Archetype	#populations	Total abundance	Relative abundance	Dispersal	Complete?
Arch2_1	2	2	7500even	50:50	0	
Arch2_2	2	2	7500even	50:50	5x10-6	X
Arch2_3	2	2	7500even	50:50	5x10-5	X
Arch2_4	2	2	7500even	50:50	5x10-4	X
Arch2_5	2	2	7500even	50:50	5x10-3	X
Arch2_6	2	2	15000even	50:50	0	
Arch2_7	2	2	15000even	50:50	5x10-6	X
Arch2_8	2	2	15000even	50:50	5x10-5	X
Arch2_9	2	2	15000even	50:50	5x10-4	X
Arch2_10	2	2	15000even	50:50	5x10-3	X
Arch2_16	2	2	7500uneven	90:10	0	
Arch2_17	2	2	7500uneven	90:10	5x10-6	X
Arch2_18	2	2	7500uneven	90:10	5x10-5	X
Arch2_19	2	2	7500uneven	90:10	5x10-4	X
Arch2_20	2	2	7500uneven	90:10	5x10-3	X
Arch2_31	2	3	7500even	33:33:33	0	
Arch2_32	2	3	7500even	33:33:33	5x10-6	X
Arch2_33	2	3	7500even	33:33:33	5x10-5	X
Arch2_34	2	3	7500even	33:33:33	5x10-4	X
Arch2_35	2	3	7500even	33:33:33	5x10-3	X
Arch2_36	2	3	15000even	33:33:33	0	
Arch2_37	2	3	15000even	33:33:33	5x10-6	X
Arch2_38	2	3	15000even	33:33:33	5x10-5	X
Arch2_39	2	3	15000even	33:33:33	5x10-4	X
Arch2_40	2	3	15000even	33:33:33	5x10-3	X
Arch2_46	2	3	7500uneven	45:45:10	0	
Arch2_47	2	3	7500uneven	45:45:10	5x10-6	X
Arch2_48	2	3	7500uneven	45:45:10	5x10-5	X
Arch2_49	2	3	7500uneven	45:45:10	5x10-4	X
Arch2_50	2	3	7500uneven	45:45:10	5x10-3	X
Arch2_61	2	2	2500even	50:50	0	
Arch2_62	2	2	2500even	50:50	5x10-6	X
Arch2_63	2	2	2500even	50:50	5x10-5	X
Arch2_64	2	2	2500even	50:50	5x10-4	X
Arch2_65	2	2	2500even	50:50	5x10-3	X
Arch2_66	2	2	2500uneven	90:10	0	
Arch2_67	2	2	2500uneven	90:10	5x10-6	X
Arch2_68	2	2	2500uneven	90:10	5x10-5	X
Arch2_69	2	2	2500uneven	90:10	5x10-4	X

Table A1 (cont.). Specifications for all scenarios specified by the TOSSM steering committee. Archetype 4 is not included in the list, as it can be generated by simply sampling from Archetype 2 simulations. 'Total abundance' refers to the combined abundance of all populations in the dataset. 'Relative abundance' indicates the percentage of the total abundance that is in each of the populations. The last column indicates whether a dataset has been generated as of 1 August, 2007. Scenario numbers are not continuous because some scenarios originally specified have since been eliminated.

Scenario name	Archetype	#populations	Total abundance	Relative abundance	Dispersal	Complete?
Arch2_70	2	2	2500uneven	90:10	5x10-3	X
Arch2_71	2	3	2500even	33:33:33	0	
Arch2_72	2	3	2500even	33:33:33	5x10-6	X
Arch2_73	2	3	2500even	33:33:33	5x10-5	X
Arch2_74	2	3	2500even	33:33:33	5x10-4	X
Arch2_75	2	3	2500even	33:33:33	5x10-3	X
Arch2_76	2	3	2500uneven	45:45:10	0	
Arch2_77	2	3	2500uneven	45:45:10	5x10-6	X
Arch2_78	2	3	2500uneven	45:45:10	5x10-5	X
Arch2_79	2	3	2500uneven	45:45:10	5x10-4	X
	2	3	2500uneven	45:45:10	5x10-3	X

Scenario name	Archetype	#populations	Total abundance	Relative abundance	Dispersal	Complete?
Arch3_1	3	2	7500	n/a	5x10-6	
Arch3_2	3	2	7500	n/a	5x10-5	
Arch3_3	3	2	7500	n/a	5x10-4	
Arch3_4	3	2	7500	n/a	5x10-3	
Arch3_5	3	2	15000	n/a	5x10-6	
Arch3_6	3	2	15000	n/a	5x10-5	
Arch3_7	3	2	15000	n/a	5x10-4	
Arch3_8	3	2	15000	n/a	5x10-3	

Scenario name	Archetype	#populations	Total abundance	Relative abundance	Dispersal	Complete?
Arch5_1	5	2	7500	50:50	0	
Arch5_2	5	2	7500	50:50	5x10-6	
Arch5_3	5	2	7500	50:50	5x10-5	
Arch5_4	5	2	7500	50:50	5x10-4	
Arch5_5	5	2	7500	50:50	5x10-3	
Arch5_6	5	2	15000	50:50	0	
Arch5_7	5	2	15000	50:50	5x10-6	
Arch5_8	5	2	15000	50:50	5x10-5	
Arch5_9	5	2	15000	50:50	5x10-4	
Arch5_10	5	2	15000	50:50	5x10-3	