Updates to HISEA program for estimation of mixed stock composition

NOTE: This program is designed for Windows PC use, and is not supported for use on other platforms.

There are some changes to the attached documentation. In particular, the names of the required input files have been changed, and a different format is required in the data files.

- The executable is `hisea.exe`
- The control file must be called `hisea.ctl`
- The baseline file must be called `hisea.std`
- The mixed fishery file (if used in analysis mode) must be called `hisea.mix`
- The stocks in `hisea.std` must be separated by a line with non-numeric input. Both data files (`hisea.std` and `hisea.mix`) must end with two such lines. (It is best that these lines do not begin with F or T, since these can be interpreted as logical, and be converted to numeric.)
- In a bootstrap or simulation, for each iteration, the five-different estimates of stock composition are written to text files. These files are named hisea1.txt to hisea5.txt.

The limitations on dimensions have also been increased. In `hisea.exe` these are now:

- No. of stocks ≤ 20
- No. of variables ≤ 50
- No. of observations in baseline (.std) file $\leq 10\ 000$
- No. of observations in mixed (.mix) file $\leq 10\ 000$
- No. of simulations or bootstraps $\leq 1\ 000\ 000$

Canadian Technical Report of
Fisheries and Aquatic Sciences 1753

September 1990

A VERSATILE COMPUTER PROGRAM FOR MIXED STOCK FISHERY COMPOSITION ESTIMATION

by

R. B. Millar

Science Branch
Department of Fisheries and Oceans
P.O Box 5667
St John's, Newfoundland A1C 5X1
Canada

ABSTRACT

Millar, R. B. 1990. A versatile computer program for mixed stock fishery composition estimation. Can. Tech. Rep. Fish. Aquat. Sci. 1753: iii + 29 p.

This technical report describes a FORTRAN program (HISEA.FOR) to be used in the study of mixed stock fishery composition estimation. Program HISEA is a versatile research tool capable of performing analyses, The simplest use of HISEA is to perform a composition analysis. The program can then be run in bootstrap mode to provide a non-parametric estimate of the reliability of the estimated compositions. In simulation mode, HISEA can quantify the effect of changing sample sizes; of adding or removing variables; and of adding or removing stocks. Simulation also identifies potential problems (such as high bias and variance in the composition estimators) that may arise due to different stocks being too alike.

RÉSUMÉ

Millar, R. B. 1990. A versatile computer program for mixed stock fishery composition estimation. Can. Tech. Rep. Fish. Aquat. Sci. 1753: iii + 29 p.

Le présent rapport technique décrit l'application d'un programme FORTRAN (HISEA.FOR) à l'étude des estimations de composition des stocks divers de poisson. Le programme HISEA est un outil de recherche à usages multiples qui peut effectuer des analyses. L'emploi le plus simple de HISEA consiste à réaliser une analyse de composition. Le programme peut alors être exécuté en mode bootstrap pour fournir une estimation non paramétrique de la fiabilité des compositions estimées. En mode de simulation, HISEA peut quantifier les résultats de changements de tailles d'échantillons, de l'ajout ou du retrait de variables et de l'ajout ou du retrait de stocks. La même simulation identifiera les problèmes potentials (comme un biais élevé et une variance des estimateurs de composition) que pourrait provoquer une trop grande ressemblance entre stocks différents.

Cat. No. Fs 97-6/1753E ISSN 0706-6457

Correct citation for this publication:

Millar, R. B. 1990. A versatile computer program for mixed stock fishery composition estimation. Can. Tech. Rep. Fish. Aquat. Sci 1753: iii + 29 p.

INTRODUCTION

The motivation for creation of the simulation/bootstrap/analysis program HISEA was to perform comparison studies of the relative performance of several different estimators of mixed fishery composition. The output of HISEA (Appendix 3) includes five different estimates. These estimators are listed below - for more detail the reader is referred to Millar (in press). The notation and terminology used throughout this technical report is consistent with that of Millar (in press).

Using HISEA, Millar (in press) found the direct maximum likelihood estimator Θ_5 to be superior and encourages the use of it alone. However, since HISEA is a research program, the other four estimators continue to be calculated. When performing long simulation or bootstrap runs one might consider revising HISEA to calculate only Θ_5 if speed of execution becomes a problem.

The original version of HISEA was written in 1987 for the University of Washington's CDC Cyber mainframe computer. In 1990 it was revised to run under Digital's VMS operating system (Appendix 4).

PROGRAM SUMMARY

ESTIMATORS

The five estimators calculated by HISEA are

Θ_1 : Raw classification proportions
Θ_2 : Cook and Lord corrected classification estimator
Θ_3 : Cook's constrained corrected classification estimator
Θ_4 : Maximum classification-likelihood estimator
Θ_5 : Direct maximum likelihood estimator

The first four estimators are based on a classification step. Inadequacies in estimator Θ_1 (it is not a statistically consistent estimator) prompted the development of estimator Θ_2 by Worlund and Fredin (1962), Cook and Lord (1978), Pella and Robertson (1979) and Cook (1982). Cook (1983) presents Θ_3 . Millar (1987) focuses on Θ_5 but also discovered its relationship with Θ_2 and Θ_3 , and proposes Θ_4 as an alternative to Θ_2 and Θ_3 . Θ_5 is used by Milner (1981,1983) and Fournier (1984).

Estimator Θ_1 is not consistent, and estimator Θ_2 is not well defined since it can produce negative estimates of composition. Thus, of the five estimators, only Θ_3 , Θ_4 or Θ_5 are statistically valid. Millar (in press) recommends using Θ_5 since it has better performance (lower mean squared error) than Θ_3 and Θ_4 .

ASSUMPTIONS

Program HISEA assumes that the variables used are approximately distributed as multi-variate normal and that they have a covariance matrix that is common to all stocks. Variables taken from scales (circuli distances or counts, etc) will usually satisfy this assumption. Researchers using other types of variables (genotypes, etc) will have to modify HISEA accordingly (see below).

Array X contains likelihood values, that is, $X(i, j)$ is the likelihood of observing the measurements made on fish i given that it is from stock j. Under the assumption of multivariate normality, calculation of the log-likelihood is defined by a linear combination of the measured variables (the linear discriminant function). These coefficients are calculated in subroutine LDF and the likelihoods $X(i, j)$ are calculated in subroutine CLASFY (this subroutine then classifies the fish on the basis of these likelihoods - see Millar (in press)).

A researcher wishing to use variables that are not multi-variate normal, or do not satisfy the constant covariance matrix assumption, will have to modify subroutines LDF and CLASFY. The modification requires only that the likelihood values $X(i, j)$ are calculated according to the assumptions made concerning the distribution of the measured variables.

IMPLEMENTATION

Estimation of the classification matrix

Program HISEA estimates the classification matrix Φ by classifying the baseline data, see Millar (in press). (Note that when doing simulation or bootstrap runs without resampling from the given baseline, then the classification matrix is fixed throughout the run. If resampling the baseline then the classification matrix will most likely be different for every bootstrap or simulation in the run.) Some authors (e.g., Pella and Robertson 1979) split the baseline data in two, using one half to devise the classification rule and estimate Φ by classifying the second half. A third possibility is to use a leave-one-approach. The difference in these methods will be negligible in practice, provided the baseline is of a reasonable size, say, ≥ 100 fish in total.

The estimator of Φ used by HISEA is the simplest and is also the one used by the SAS discrimination procedure DISCRIM. In addition, when simulating without resampling the baseline, the estimated Φ is, in fact, the true classification matrix since the baseline defines the population being simulated (Millar, in press).

Maximization of Likelihood

Calculation of estimators Θ_4 and Θ_5 is a constrained non-linear maximization problem. There are several ways to perform this maximization – fortunately, the choice of method does *not* affect the estimators. One simply requires an algorithm that can be relied upon to converge to the maximum of the likelihood function.

Program HISEA uses the expectation-maximization (EM) algorithm (Dempster et al. 1977) within subroutine EM. The advantage of the EM algorithm is that, in this application, it is exceptionally easy to program and it can be written in a single line (Milner et al. 1983; Millar 1987). The algorithm is extremely robust and it never fails to converge to the maximum. The EM algorithm used by HISEA has been accelerated (subroutine ACCEL) and this speeds up the algorithm when the estimates from successive iterations are close to converging. The maximum number of iterations has been set to 100 (integer variable IMAX) and the algorithm is deemed to have converged if the maximum absolute difference between successive estimates of composition is less than 10^{-6} (real variable TOL).

INSTRUCTIONS FOR RUNNING HISEA

The program requires 3 input files for analysis or bootstrap runs, the baseline data (data on fish of known origin), a control file with parameters and options, and a file with the mixed sample (unknown origin) data. A simulation run does not need the later file since it generates its own mixed sample data. The default names for the input files are FOR007.DAT, FOR008.DAT and FOR009.DAT respectively. Examples of FOR007.DAT and FOR008.DAT are given in Appendices 1 and 2 respectively. The output produced by running HISEA with these input files is given in Appendix 3.

Below is an example of the control file FOR008.DAT, followed by an explanation of the records on each line.

```
Line 1   '1985 Atlantic salmon trial run'
Line 2   'SIMULATION'
Line 3   4   'Europe'   'Canada'   'USA'   'Greenlan'
Line 4   9
Line 5   'STD'   'Y'
Line 6   500   500   500   500
Line 7   'MIX'   Y'
Line 8   84357
Line 9   10
Line 10   100   0.00   0.00   0.00   1.00
```

Line 1: Title; ≤ 80 characters.

Line 2: Type of run (Simulation, analysis, or bootstrap).

Line 3: Number and names of stock groups; number must be ≤ 8; only the first 8 characters of the names are used.

Line 4: Number of variables (≤ 16).

Line 5: Whether or not to resample the baseline. This line must read either 'STD' 'N' or 'STD' 'Y', corresponding to no resampling and resampling respectively. (Here STD is an abbreviation for "standard", which is used instead of "baseline" by some researchers.)

Line 6: Size desired for resampled baselines; can be any size (sampling with replacement); must be positive; must include this line even if not resampling (option 'N' on line 5).

Line 7: Whether or not to resample/simulate the mixed sample. This line must read either 'MIX' 'N' or 'MIX' 'Y' . If simulating, the mixed sample is generated from the baseline data (FOR007.DAT) and 'MIX' 'Y' must be specified.

Line 8: Seed for random number generator; must be a (large) positive integer.

Line 9: Number of simulations desired; must be between 1 and 1000.

Line 10: Size of mixed sample to be taken from a simulated mixed fishery population with composition given by the remaining values on this line (this line read only for simulations).

All lines are freefield format. Only the first four lines are read for an Analysis run.

The baseline and mixed sample files are also freefield format and consist of variable values only, with no header or other identifier information on the line. The mixed sample file FOR009.DAT contains variable values for fish from a sample of the mixed population. This file is not required if HISEA is used in simulation mode.

In the version of HISEA presented here, the baseline and mixed sample files must not exceed 1400 lines each.

The baseline file (FOR007.DATA) contains the baseline data for all of the stock groups, linked end-to-end and separated by the end-of-file marker. (On the Northwest Atlantic Fisheries Centre's VAX 6310 this is the control Z character and it can be inserted with the edt editor by hitting the `<PF1>` key followed by the 3 on the extended keyboard. At the command line prompt type 26 (from the regular keyboard) and hit the `<Do>` key. If done correctly the character will appear as `^Z` and will be inserted at the current position of the cursor.) Each line in FOR007.DAT represents the measurements made on a fish of known origin. The stocks are assumed to be in the order as given in line 3 above. The program automatically determines the size of the baseline (i.e., number of fish) for each stock by using the `^Z` markers.

PORTABILITY

Program HISEA has also been ported to SUN Microsystems's workstations. The intrinsic functions that appear not be compatible across all implementations of FORTRAN are DATE and TIME (subroutine READ8) and the random number generator RAN (subroutine ORDVEC). The code in subroutine READIN may need to be modified if the end-of-file marker cannot be inserted in file FOR007.DAT.

APPENDIX 1

This is a sample baseline data file (FOR007.DAT) with 32 and 24 fish respectively. The data are circuli counts and are a subset of a 1985 baseline for European and North American atlantic salmon.

```
26  2
22  7
24  7
27  5
26  4
26  2
33  2
29  6
25  4
21  5
33  4
24  5
32  3
29  5
29  5
23  5
23  4
28  2
27  5
25  4
30  3
25  2
21  4
30  3
26  3
29  2
29  2
27  3
34  2
29  5
29  6
32  3
 2
18  10
32  3
24  2
26  5
27  7
24  3
27  2
30  5
19  9
12  4
20  7
14  4
16  4
19  3
13  9
23  6
21  5
24  5
19  8
25  8
26  5
22  13
28  7
23  6
```

APPENDIX 2

A sample control file (FOR008.DAT) for performing a simulation using the baseline data from Appendix 1. Note that the resampled baselines are specified to be size 50 and that the mixed sample size is 200. HISEA will perform 100 simulations from a hypothetical mixed fishery with equal contribution of European and North American fish.

```
'1985 Atlantic salmon trial run'
'SIMULATION'
2 'Europe' 'Nth America'
2
'STD' 'Y'
50 50
'MIX' 'Y'
123456
100
200 0.9 0.1
```

APPENDIX 3

Listing from simulation run using the input files listed in Appendices 1 and 2. The maximum classification-likelihood estimator and direct maximum likelihood estimator are labelled the "Millar constrained" and "maximimum likelihood" estimators respectively. This run took 8 seconds of CPU time on a VAX 6310.

```
$run hisea

PROGRAM  HISEA.......EXECUTION DATE:  9-APR-90     14:04:30

1985 Atlantic salmon trial run


FUNCTION OF THIS RUN IS.........SIMULATION
#STOCKS IN THE MODEL..... 2
THE STOCKS ARE..................Europe   Nth Amer
#VARIABLES USED.................. 2

STANDARD BEING RESAMPLED?.......Y
RESAMPLED STANDARD SIZES........    50    50
MIXTURE BEING SIMULATED?........Y

RANDOM NUMBER GENERATOR SEED....  123456
NUMBER OF RUNS REQUESTED?.......  100
SIZE OF SIMULATED MIXTURE.......  200
ACTUAL COMPOSITION IS...........   0.900   0.100

======================================================================

                        MEAN AND STD DEV SUMMARY OF VARIABLES

VAR       Europe          Nth Amer

 1         27.281          22.167
           (3.49)          (5.25)

 2         3.8750          5.8333
           (1.52)          (2.71)

======================================================================


THE SIZES OF THE STANDARDS ARE     32    24

TABLE OF COMPOSITION ESTIMATE MEANS.  NUMBER OF RUNS =  100

            RAW      COOK & LORD      COOK        MILLAR      MAXIMUM
                                   CONSTRAINED  CONSTRAINED  LIKELIHOOD

Europe    0.7420       0.9172       0.8914       0.8913       0.9470
Nth Amer  0.2580       0.0828       0.1086       0.1087       0.0530
```

TABLE OF COMPOSITION ESTIMATE STANDARD DEVIATIONS OVER THE 100 RUNS

	RAW	COOK & LORD	COOK CONSTRAINED	MILLAR CONSTRAINED	MAXIMUM LIKELIHOOD
Europe	0.0359	0.1406	0.0972	0.0971	0.0489
Nth Amer	0.0359	0.1406	0.0972	0.0971	0.0489

TABLE OF SQRT OF MEAN SQUARED ERRORS

	RAW	COOK & LORD	COOK CONSTRAINED	MILLAR CONSTRAINED	MAXIMUM LIKELIHOOD
Europe	0.1620	0.1416	0.0976	0.0975	0.0678
Nth Amer	0.1620	0.1416	0.0976	0.0975	0.0678

THE COVARIANCE MATRIX OF THE 100 MAXIMUM LIKELIHOOD COMPOSITION ESTIMATES IS:

(POPULATIONS 1- 2 * 1- 2)

	1	2
1	0.2395E-02	
2	-0.2394E-02	0.2394E-02

THE CORRESPONDING CORRELATION MATRIX IS:

(POPULATIONS 1- 2 * 1- 2)

	1	2
1	0.1000E+01	
2	-0.9999E+00	0.1000E+01

FORTRAN STOP
KIWI job terminated at 9-APR-1990 14:04:39.39

Accounting information:

Buffered I/O count:	83	Peak working set size:	490
Direct I/O count:	64	Peak page file size:	3003
Page faults:	796	Mounted volumes:	0
Charged CPU time:	0 00:00:08.02	Elapsed time:	0 00:00:11.96