Not to be cited without permission of the author[1]

Ne pas citer sans autorisation des auteur[1]

Canadian Atlantic Fisheries Scientific Advisory Committee

Comité scientifique consultatif des pêches canadiennes dans l'Atlantique

CAFSAC Research Document 88/ 29

CSCPCA Document de recherche 88/29

An adaptive framework for the estimation of population size

by

Stratis Gavaris
Marine Fish Division
Department of Fisheries and Oceans
Biological Station
St. Andrews, New Brunswick EOG 2X0

[1]This series documents the scientific basis for fisheries management advice in Atlantic Canada. As such, it addresses the issues of the day in the time frames required and the Research Documents it contains are not intended as definitive statements on the subjects addressed but rather as progress reports on ongoing investigations.

Research Documents are produced in the official language in which they are provided to the Secretariat by the author.

[1]Cette série documente les bases scientifiques des conseils de gestion des pêches sur la côte atlantique du Canada. Comme telle, elle couvre les problèmes actuels selon les échéanciers voulus et les Documents de recherche qu'elle contient ne doivent pas être considérés comme des énoncés finals sur les sujets traités mais plutôt comme des rapports d'étape sur les études en cours.

Les Documents de recherche sont publiés dans la langue officielle utilisée par les auteurs dans le manuscrit envoyé au secrétariat.

ABSTRACT

An adaptive framework for the estimation of population size, based on minimizing the discrepancy between observations of variables and the values of those variables predicted as functions of population parameters, provides a statistical basis for this type of problem. The flexibility in the types of data and relationships which may be employed is considered essential in order to handle the wide range of situations encountered in stock assessments. Application of the framework is demonstrated with two diverse hypothetical situations. An implementation using a Marquardt algorithm for minimization of the objective function has been developed using the APL programming language.

RÉSUMÉ

Un outil de travail permet d'adapter des méthodes statistiques à l'estimation de l'effectif, afin de réduire les différences entre l'observation des variables et leur valeur, à partir d'une prévision basée sur les paramètres des la population. La souplesse vis-à-vis des données et des relations utilisables est considérée essentielle face aux diverses situations qui se présentent dans l'évaluation des populations. Deux situations hypothétiques différentes illustrent l'application de l'outil de travail. On a mis un point un programme en langage de programmation APL qui fait appel à un algorithme Marquardt pour minimiser la fonction objective.

INTRODUCTION

The status of groundfish stocks in the northwest Atlantic have typically been assessed in recent years using a procedure which has become commonly referred to as "calibration" of the sequential population analysis (Anon. 1987a; Anon. 1987b). Briefly, a population numbers by age matrix (table) is generated from the catch-at-age matrix using a sequential population analysis, such as cohort (Pope 1972). To generate the population matrix, either the population size or fishing mortality for the most recent year and the oldest age group must be assumed. The relationships between the population matrix and abundance indices for selected age groups are then derived by ordinary least squares. Several values are tried for the population size or fishing mortality in the most recent year and those which result in the "best" relationships are considered the most appropriate estimates. Age groups are generally processed sequentially from oldest to youngest.

This procedure has been criticized because it does not properly account for the sampling errors of measured variables or of model errors due to departures from hypothesized relationships. Another shortcoming relates to the inappropriate use of diagnostics intended for investigation of "goodness of fit," such as residuals, R^2 and significance of intercepts, as a means of selecting the "best" estimate. Therefore, this procedure is not founded on established statistical principles and the properties of this "best" estimator are not defined. Finally, the values in a population matrix from sequential population analysis have an interdependence along cohorts. Treating age groups sequentially may give results which are unduly influenced by the relationships at older ages and which do not account for effects at younger ages.

In recent years, several workers have considered appropriate statistical formulations of this problem (Collie and Sissenwine 1983; Deriso et al. 1985; Fournier and Archibald 1982). The formulations differed primarily in the details of the structural relationships and the treatment of the errors. This paper describes a framework which can be employed in describing the various formulations that this problem can take. An algorithm for implementing the framework and a brief description of the numerical methods employed is included.

MODEL

The basic framework is a mathematical expression for the application of a common statistical technique, least squares, in order to determine the most appropriate estimate of a population matrix. The discrepancy between observations of variables and the values of those variables predicted as functions of the population matrix are employed for this determination. That is,

$$\min\sum_i (W_i(O_i - f(\Pi,\Omega)))^2 \tag{1}$$

where Π = population matrix;
O_i = observed variable set i;
W_i = weight for observed variable set i; and
Ω = other parameters which may be required.

Both O_i and W_i may be matrices or vectors (series). The weighting factors, W_i, are needed to accommodate differences in the reliability of the elements within an observed variable set as well as differences in reliability between variable sets. In the absence of measures of precision for use as weights, transformations may be employed in attempting to stabilize variance. The summation is taken over all sets i as indicated, as well as within each set.

The framework is adaptive in the sense that any observed variable which is a function of the population matrix can be accommodated by equation 1. Further, various formulations of the structural relationships and statistical error models which link these observed variables with the population matrix may be invoked. This flexibility is considered essential given the wide range of situations encountered in stock assessments. Common statistical diagnostics, (residual plots, probability plots, standard error of parameters, correlation matrices of parameters being estimated) are used to select from among the formulations, those which are most suitable for the particular conditions experienced. To elucidate the basic framework and to demonstrate the flexibility in the types of relationships which may be employed, two hypothetical scenarios are described.

Scenario A

The commercial catch has been sampled using a double sampling design and the estimated catch at age, C_{ay}, is available with the associated standard error, ${}_cS_{ay}$. It is known that age determinations for older ages are variable, therefore ages 1-5 are treated individually while ages 6 and older are aggregated. There is no reliable information on effort from the commercial fishery. A research survey index of abundance at age, I_{ay}, is available for all ages. The survey was conducted at the beginning of the year using a stratified random sampling design and the appropriate standard error for the index, ${}_IS_{ay}$, has been derived. There are no other relevant observed variables. The expression to be minimized is:

$$\sum_{a=1}^{6+}\sum_{y=1}^{20}\left(\frac{1}{{}_cS_{ay}}(C_{ay}-\hat{C}_{ay})\right)^2 + \sum_{a=1}^{6+}\sum_{y=1}^{21}\left(\frac{1}{{}_IS_{ay}}(I_{ay}-\hat{I}_{ay})\right)^2 \tag{2}$$

where a = index for age;
y = index for year.

The beginning of the year survey is available at the time the assessment is done, allowing for 21 yr of the survey index to be used while only 20 yr of catch at age data are available. To ensure that population size decreases along cohorts with time, the parameter set Π is replaced by Θ, an estimate of year-class size for each cohort, and Φ, the fishing mortality matrix. The associated population matrix can then be calculated using the relationship:

$$P_{ay} = P_{a+1\ y+1}\exp[\Phi_{a+1\ y+1} + M] \tag{3}$$

where natural mortality, M, is assumed to be constant. The appropriate cohort year-class size, Θ, is substituted for P as needed. The predicted catch, $\hat{C}$, can then be obtained using the catch equation:

$$\hat{C}_{ay} = (\phi_{ay}P_{ay}(1 - \exp[-\Phi_{ay} + M]))/(\Phi_{ay} + M) \tag{4}$$

A linear relationship through the origin is assumed between the abundance index and population size. Therefore the predicted index, I, is obtained from:

$$\hat{I}_{ay} = \kappa_a P_{ay} \tag{5}$$

where κ_a = calibration coefficient for age a. The parameter set Ω (of Equation 1) consists of only κ_a in this scenario. Equations 2-5 can be used to solve for the least squares estimates of Θ, Φ and κ.

Scenario B

The commercial catch has been sampled as in scenario A above; however, the errors in the estimates of catch at age are considered negligible. A combined catch rate series, U_y, has been derived with a multiplicative model and its associated standard error was ${}_uS_y$. There are two research survey abundance indices, I_1 and I_2, and their standard errors, ${}_IS_1$ and ${}_IS_2$, were computed based on the survey design employed. Survey index I_2 is considered a recruitment index, suitable for the first two ages only, and is only available for the most recent 6 yr. Both surveys are related to beginning of year population. The expression to be minimized is therefore:

$$\sum_{a=1}^{10}\sum_{y=1}^{21}\left(\frac{1}{{}_{I_1}S_{ay}}(I_{1ay}-\hat{I}_{1ay})\right)^2 + \sum_{a=1}^{2}\sum_{y=16}^{21}\left(\frac{1}{{}_{I_2}S_{ay}}(I_{2ay}-\hat{I}_{2ay})\right)^2 + \sum_{y=1}^{20}\left(\frac{1}{{}_uS_y}(U_y-\hat{U}_y)\right)^2 \tag{6}$$

Since errors in the catch at age are considered negligible, the parameter set Π is reduced to Θ, the year-class size for each cohort i.e. only one cell of each cohort in the population matrix needs to be designated as a parameter. The last year and oldest age are used as the designated cell of the year-class size for each cohort. The population matrix can then be derived using:

$$P_{ay} = C_{ay}\exp[M/2] + P_{a+1\ y+1}\exp[M] \tag{7}$$

where the appropriate cohort year-class size, Θ, is substituted for P as needed. Linear relationships are assumed for both survey indices. Intercepts are accepted for survey index I_2 even though a mechanism to generate such a relationship has not been established. Therefore:

$$\hat{I}_{1ay} = \kappa_{1a}P_{ay} \tag{8}$$

and

$$\hat{I}_{2ay} = \kappa'_{2a} + \kappa_{2a}P_{ay} \tag{9}$$

A fishing mortality matrix is calculated from:

$$F_{ay} = \ln(P_{ay}/P_{a+1\ y+1}) - M \tag{10}$$

The partial fishing mortality matrix for the otter trawl fleet was obtained as:

$$_TF_{ay} = F_{ay}(_TC_{ay}/C_{ay}) \tag{11}$$

The annual fully recruited fishing mortality for trawlers was derived using:

$$_TF'_y = \sum_{a=5}^{10} (P_{ay}\ {_TF_{ay}})/\sum_{a=5}^{10} (P_{ay}) \tag{12}$$

The annual partial recruitment for the trawler fleet is then obtained:

$$_TPR_y = {_TF_{ay}}/{_TF'_y} \tag{13}$$

and used to calculate the average annual exploitable biomass:

$$_{T}\bar{B}_{y}' = \bar{W}_{ay}(P_{ay}((1-\exp[-F_{ay} + M]/F_{ay} + M)) \; _{T}PR_{y} \tag{14}$$

A linear relationship through the origin is hypothesized for the otter trawl catch rate and the exploitable biomass:

$$\hat{U}_{y} = \kappa_{3} \; _{T}\bar{B}_{y}' \tag{15}$$

We now have the quantities required for minimization of expression 6.

IMPLEMENTATION

Implementation of an algorithm to solve this model may be decomposed into two parts, a routine for minimization of an objective function and the specification of an objective function. Such a partitioning allows the use of readily available, general purpose software for the minimization module.

In its present form, a Marquardt algorithm, coded in the programming language APL, is used to minimize the objective function. This method has been used successfully for a broad range of non-linear problems. The objective function, which must be defined by the user, is expected to produce a vector of residuals to be minimized, one for each element of observed variables. The residuals may be weighted if this is appropriate.

The Marquardt algorithm is described in detail by Bard (1974, p. 94). An iterative scheme is used to improve the initial guess, provided by the user, of the parameters being estimated. This method is a modification of the Newton-Raphson method, designed to overcome the problem of indefiniteness of the Hessian matrix (the matrix of second partial derivatives). To improve the performance, the facility to define constraints on the parameters via a penalty function (Bard 1974, p. 141) has been included. The constraints are removed for the final iterations. Also, to simplify matters, the partial derivatives are obtained numerically.

The objective function component is accommodated by providing various general purpose modules such as a sequential population function, a function to calculate residuals for an abundance index, a function to calculate residuals for the catch at age, etc. The user can then define an appropriate objective function for a specific formulation by calling relevant modules. Further, modules are accessible for modification as required. This tactic was taken to allow for the utmost in flexibility when specifying a formulation of the adaptive framework.

SUMMARY

The development and implementation of an adaptive framework was based on the observation that an assessment technique is composed of two parts, a method of estimation and the definition of a model. For the method of estimation, a statistically established method for solving the problem of estimating the parameters of a model was selected. With respect to the definition of a model, it was recognized that the variety of situations encountered would require considerable flexibility. The philosophy of an adaptive approach was taken rather than attempting to identify a model which would perform "well" in "typical" situations. Common diagnostic tools, residual analysis, evaluation of variance/co-variance of parameters, etc., are employed to determine the appropriate formulation used in the adaptive framework.

In addition to the advantages offered by the flexibility of the technique, the parameter estimates obtained have certain desirable statistical properties, given the correct model specification, and their reliability can be evaluated using the covariance matrix. A discussion of the properties of least squares estimators is given by Bard (1974). A feature of the technique which may be perceived as a limitation in some instances, is the requirement to be fairly familiar with the fishery and the data for the stock being considered. It is not designed to encourage automatic processing of an arbitrary data set by an unknowledgeable user.

ACKNOWLEDGEMENTS

Many participants at meetings of the Groundfish Subcommittee and the Statistics, Sampling and Survey Subcommittee of CAFSAC have contributed to the development of these concepts and their interest was greatly appreciated. In addition, I am thankful to J. Neilson, R. Stephenson and K. Zwanenburg for reviewing an earlier draft.

REFERENCES

Anon. 1987a. Canadian Atlantic Fisheries Scientific Advisory Committee Annual Report, Vol. 9. Dept. Fisheries and Oceans, Dartmouth, Nova Scotia, Canada.

Anon. 1987b. Northwest Atlantic Fisheries Organization Scientific Council Reports 1987. NAFO, Dartmouth, Nova Scotia, Canada.

Bard, Y. 1974. Nonlinear parameter estimation. Academic Press, New York, NY. 341 p.

Collie, J. S., and M. P. Sissenwine. 1983. Estimating population size from relative abundance data measured with error. Can. J. Fish. Aquat. Sci. 40: 1871-1879.

Deriso, R. B., T. J. Quinn II, and P. R. Neal. 1985. Catch at age analysis with auxilliary information. Can. J. Fish. Aquat. Sci. 42: 815-824.

Fournier, D., and C. P. Archibald. 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39: 1195-1207.

Pope, J. G. 1972. An investigation of the accuracy of virtual population analysis using cohort analysis. Res. Bull. Int. Comm. Northw. Atl. Fish. 9: 65-74.

ANNEX

Listing of functions in the APL workspace NLLS which implements a Marquardt algorithm for minimization of an objective function.

```
     ∇ R←DIFF∆OBJ;DELTA;I;TPAR
[1]    ⍝ CALCULATES ONE SIDED DIFFERENCE OF OBJECTIVE FUNCTION
[2]    I←1
[3]    R←(N,0)⍴1
[4]    DELTA←(0.01×par)+0.01×par=0
[5]    L1:TPAR←((I-1)↑par),(par[I]+DELTA[I]),I↓par
[6]      R←R,(e-OBJ∆FN TPAR)÷DELTA[I]
[7]      →L1×P≥I←I+1
     ∇
     ∇ R←DIFF∆PNLTY;I;R1;DELTA;TPAR;fpnlty;bpnlty
[1]    ⍝ CALCULATES FIRST AND SECOND DIFFERENCES OF PENALTY FUNCTION
[2]    I←1
[3]    R←2 0⍴0
[4]    DELTA←(0.01×PAR)+0.01×PAR=0
[5]    L1:TPAR←((I-1)↑PAR),(PAR[I]+DELTA[I]),I↓PAR
[6]      R1←(pnlty-fpnlty←alpha PNLTY∆FN TPAR)÷DELTA[I]
[7]      TPAR←((I-1)↑PAR),(PAR[I]-DELTA[I]),I↓PAR
[8]      bpnlty←alpha PNLTY∆FN TPAR
[9]      R←R,,R1,(fpnlty+bpnlty-2×pnlty)÷DELTA[I]
[10]     →L1×P≥I←I+1
     ∇
     ∇ R←FRGN∆FN A
[1]    ⍝ THIS FUNCTION SHOULD RETURN A 1 IF THE PARAMETERS
[2]    ⍝ ARE IN THE FEASIBLE REGION AND 0 OTHERWISE
[3]    R←1             ⍝ DEFAULT RETURNS 1
     ∇
```

```
    ∇ NLLS;BOOL;J;DIAG;Q;LAMBDA;HESS;N;P;PAR;RSS;de;CAUSE;I;V;NPHI;PHI;pnlty;d
      pnlty;SHESS;NORM
[1]   ⍝ NON-LINEAR LEAST SQUARES USING MARQUARDT ALGORITHM
[2]    P←⍴par←PAR←,initial
[3]    RSS←e+.×e←OBJ∆FN PAR                  ⍝ RESIDUAL SUM OF SQUARES
[4]    N←⍴,e
[5]    pnlty←alpha PNLTY∆FN PAR              ⍝ PENALTY FOR CONSTRAINTS
[6]    NPHI←PHI←RSS+pnlty
[7]    LAMBDA←0.01
[8]    BOOL←(P×P)⍴1,P⍴0                      ⍝ USED TO CREATE DIAG MATRIX
[9]    con←10
[10]   PRNT
[11]   J←1
[12]  L3:→(limit<J←J+1)/L6                            ⍝MAIN LOOP
[13]    PAR←par
[14]    PHI←NPHI
[15]    de←DIFF∆OBJ
[16]    Q←2×e+.×de                                  ⍝ GRADIENT
[17]    HESS←2×XP de←↓[1]de                         ⍝ HESSIAN
[18]    dpnlty←DIFF∆PNLTY                           ⍝ DIFFERENCE FOR PENALTY
[19]    Q←Q+dpnlty[1;]
[20]    DIAG←1 1⍉HESS←HESS+(2⍴P)⍴BOOL\dpnlty[2;]
[21]    LAMBDA←0.000001⌈LAMBDA×0.01
[22]    I←1
[23]    SHESS←HESS+(2⍴P)⍴BOOL\DIAG×LAMBDA←LAMBDA×10  ⍝ MARQUARDT METHOD
[24]    NORM←(+⌿SHESS*2)*.5                           ⍝ COLUMN NORMS
[25]    SHESS←⍉↑(↓[1]SHESS)÷¨NORM                     ⍝ SCALE HESSIAN

[26]    par←PAR+V←(Q⌹SHESS)÷NORM              ⍝ STEP DIRECTION; STEP SIZE=1
[27]    →(~FRGN∆FN par)/L4
[28]    RSS←e+.×e←OBJ∆FN par
[29]    pnlty←alpha PNLTY∆FN par
[30]    →(PHI≥NPHI←RSS+pnlty)/L6
[31]    L4:LAMBDA←LAMBDA×100
[32]    L5:par←PAR+V←V×0.1*I                 ⍝INNER LOOP  REDUCE STEP SIZE
[33]      →(10<I←I+1)/L6
[34]      →(~FRGN∆FN par)/L5
[35]      RSS←e+.×e←OBJ∆FN par
[36]      pnlty←alpha PNLTY∆FN par
[37]      →(PHI≥NPHI←RSS+pnlty)/L6
[38]      →L5
[39]  L6:PRNT
[40]  msr←RSS÷N-P
[41]  →(1=^/CAUSE←(10≥I),(limit≥J),(0.001<con←(((N-P)×Q+.×V)÷P×RSS)*0.5),(0.0000
      1<|(NPHI-PHI)÷PHI),(0.00001∨.<|(par-PAR)÷0.0000000000000000001+|PAR))/L3
[42]   (~CAUSE)/[1] exit
    ∇
```

```
     ∇ R←OBJ∆FN A
[1]    ⍝ THE OBJECTIVE FUNCTION TO BE MINIMIZED
[2]    ⍝ R IS THE RESULTANT MAGNITUDE
[3]    ⍝ A IS THE PARAMETER VECTOR
[4]    'THE OBJECTIVE FUNCTION HAS NOT BEEN DEFINED'
[5]    →
    ∇
     ∇ PAR∆SE;N;P;HESS;de;NORM
[1]    'APPROXIMATE STATISTICS ASSUMING LINEARITY NEAR SOLUTION'
[2]    ' '
[3]    N←⍴,e
[4]    P←⍴,par
[5]    de←DIFF∆OBJ
[6]    HESS←2×XP de←↓[1]de
[7]    NORM←(+/HESS*2)*.5
[8]    HESS←⌹⍉↑(↓[1]HESS)÷¨NORM
[9]    HESS←2×msr×↑(↓HESS)÷¨NORM
[10]   par∆se←(1 1⍉HESS)*¯.5
[11]   corr←HESS←↑(↓[1]↑(↓HESS)×¨par∆se)×¨par∆se
[12]   par∆se←÷par∆se
[13]   'ORTHOGONALITY OFFSET.........',,'F16.6' ⎕FMT con
[14]   'MEAN SQUARE RESIDUALS .......',,'F16.6' ⎕FMT msr
[15]   ' '
[16]   '    PAR. EST.           STD. ERR.             C. V.    '
[17]   '    ---------           ---------           ----------  '
[18]   'E16.6,X3' ⎕FMT(par;par∆se;par∆se÷par)
    ∇
     ∇ R←alpha PNLTY∆FN A
[1]    ⍝ THE PENALTY FUNCTION FOR THE CONSTRAINTS
[2]    ⍝ R IS THE RESULTANT MAGNITUDE
[3]    ⍝ A IS THE VECTOR OF PARAMETERS
[4]    ⍝ alpha IS THE VECTOR OF CONSTANTS FOR THE CONSTRAINTS
[5]    R←0             ⍝ DEFAULT: NO CONSTRAINTS
    ∇
     ∇ PRNT;TMP
[1]    2 1⍴' '
[2]    TMP←3 6⍴'LAMBDARSS   NPHI  '
[3]    '10A1,E15.6' ⎕FMT (3 10↑TMP;⍎¯1↓,TMP,',')
[4]    ' '
[5]    'par'
[6]    ,'E15.6' ⎕FMT par
    ∇
     ∇ R←XP H;N;I;V
[1]    N←⍴H
[2]    R←(2⍴N)⍴I←0
[3]    V←0,⍳N
[4]    L10:→(N<I←I+1)/L99
[5]       V←1↓V
[6]       R[I;V]←+/¨H[I]×¨(I-1)↓H
[7]       R[;I]←R[I;]
[8]       →L10
[9]    L99:
    ∇
```