



INSTITUTO NACIONAL DE INVESTIGACIÓN  
Y DESARROLLO PESQUERO

# INFORME DE ASESORAMIENTO Y TRANSFERENCIA

**INIDEP**

Número

Páginas

Fecha de aprobación

Dirección

DIRECCIÓN DE PESQUERIAS DEMERSALES

Programa / Gabinete

Gabinete de Biomatemática

Actividad

## IMPLEMENTACIÓN DEL ANÁLISIS DE COHORTES CON LA ESTIMACIÓN DE PARÁMETROS POR EL MÉTODO DE MÁXIMA VEROSIMILITUD Y SIMULACIÓN DE MONTECARLO PARA REPLICACIÓN DE PARÁMETROS

Se desarrolló el programa DIAGSIM en el cual se implementa un Análisis de Cohortes para poblaciones estructuradas por edades con estimación parámetros por el método de máxima verosimilitud y una simulación de Monte Carlo para generar réplicas de los parámetros. En este trabajo se presenta el modelo, los detalles de la implementación, las instrucciones de uso y un ejemplo.

Se pretende que este programa constituya una herramienta de apoyo para los investigadores que trabajen en evaluación de recursos pesqueros.

Citar Indicando la fuente. El contenido no debe ser reproducido total o parcialmente sin la expresa conformidad del INIDEP

**SOLICITADO POR**

Institución

INIDEP

Cargo

**PREPARADO POR**

Firma:

Nombre: VERON, JUAN MANUEL

Firma:

Nombre: MENDIOLAR, MANUELA

**APROBADO POR**

Jefe de Programa / Gabinete

Director de área

Director Nacional de Investigación

Director del INIDEP

# IMPLEMENTACIÓN DEL ANÁLISIS DE COHORTES CON ESTIMACIÓN DE PARÁMETROS POR MÁXIMA VEROSIMILITUD Y SIMULACIÓN DE MONTE CARLO PARA REPLICACION DE PARAMETROS

Juan Manuel Verón y Manuela Mendiolar

Gabinete de Biomatemática. Instituto Nacional de Investigación y Desarrollo Pesquero (INIDEP),  
Paseo Victoria Ocampo N° 1 Escollera Norte, B7602HSA - Mar del Plata, Argentina.

E-mail: jveron@inidep.edu.ar; mmendiolar@inidep.edu.ar

## INTRODUCCIÓN

Tradicionalmente el Gabinete de Biomatemática colabora con diferentes Programas y Gabinetes de Investigación del INIDEP, principalmente en temas de Evaluación de Recursos Pesqueros. La colaboración se ve reflejada, entre otras actividades, en la planificación y realización de proyectos de investigación propios tendientes a mejorar las metodologías aplicadas en las evaluaciones de los recursos.

Actualmente en el Instituto la mayoría de los Programas emplea modelos estructurados por edades, para describir y predecir el impacto de la explotación en la población de peces bajo estudio (la estimación de abundancia, los estudios asociados al riesgo biológico y la recomendación de una captura máxima biológicamente aceptable para cada año son algunos ejemplos). La sobre-parametrización, inconsistencias y una inapropiada implementación del modelo son factores determinantes al momento de querer concretar lo que se desee y lograr estimaciones confiables.

Se desarrolló el programa DIAGSIM en el cual se implementa el Análisis de Cohortes (AC), modelo de dinámica poblacional estructurado por edades, con estimación parámetros por el método de máxima verosimilitud y una simulación de Monte Carlo para generar réplicas de los parámetros. El objetivo central de este trabajo es presentar el modelo, los detalles de la implementación y las instrucciones de uso. Se pretende que este programa constituya una herramienta de apoyo para los investigadores que trabajen en evaluación de recursos pesqueros.

## MODELO MATEMÁTICO

El programa que se desarrolló considera como modelo base un Análisis de Cohorte (AC), sin embargo las ideas básicas que se presentan aquí sobre la calibración son fácilmente transferibles a otros modelos estructurados por edades como el Análisis de Poblaciones Virtuales (APV), Análisis de Cohortes Modificado (ACM), Age Structured Production Model (ASPM), por dar algunos ejemplos.

### Análisis de Cohortes (AC)

El Análisis de Cohortes (AC) fue desarrollado por Pope [Pope, 1972]. Las expresiones básicas son:

$$N_{i+1,t+1} = \left( N_{i,t} e^{-\frac{M_{i,t}}{2}} - C_{i,t} \right) e^{-\frac{M_{i,t}}{2}} \quad i = 1 \dots k-2$$

$$N_{k+,t+1} = N_{k-1,t} e^{-M_{k-1,t} - F_{k-1,t}} + N_{k+,t} e^{-M_{k+,t} - F_{k+,t}}$$

$$C_{i,t} = \frac{F_{i,t}}{F_{i,t} + M_{i,t}} (1 - e^{-M_{i,t} - F_{i,t}}) N_{i,t} \quad i = 1 \dots k+$$

$$F_{i,t} = \log\left(\frac{N_{i+1,t+1}}{N_{i,t}}\right) - M_{i,t} \quad i = 1 \dots k-2$$

donde:

$N_{i,t}$  es el número de individuos de la edad  $i$  en el tiempo  $t$

$F_{i,t}$  es la tasa instantánea de mortalidad por pesca de la edad  $i$  en el tiempo  $t$

$M_{i,t}$  es la tasa instantánea de mortalidad natural de la edad  $i$  en el tiempo  $t$

$C_{i,t}$  es la captura en número de individuos de la edad  $i$  en el tiempo  $t$

$k$  es el número de clases de edad

$k+$  es el grupo plus que agrupa edades mayores o iguales que  $k$

$t_1, \dots, t_m$  es el período de evaluación

El procedimiento de cálculo puede verse en detalle en [Aubone, 2010] y para el cálculo de las tasas instantáneas de mortalidad por pesca de la edad  $k-1$  y del grupo de edades  $k+$  se emplea el algoritmo allí propuesto que permite reducir el número de parámetros a estimar.

### Parámetros del modelo

El vector de parámetros  $\theta$  del modelo está dado por

$$\theta = (F_{1,t_m}, F_{2,t_m}, \dots, F_{k-1,t_m})$$

El vector de tasas de mortalidad natural se asume fijo y de componentes iguales. Es por esto que el número de parámetros del modelo es  $k-1$ .

### Indices

Se consideró la siguiente relación entre índices de abundancia media y abundancia estimada:

$$I_t = q \overline{B}_t \varepsilon_t \quad \text{con} \quad \log(\varepsilon_t) \sim \mathcal{N}(0, \sigma_t^2), \quad q \geq 0$$

donde  $\varepsilon_t$  es el término del error en el tiempo  $t$  con  $\sigma_t^2$  su respectiva varianza y  $\overline{B}_t$  es la biomasa media en el tiempo  $t$ .

### Estimación de parámetros por el método de máxima verosimilitud

En este caso, la estimación por máxima verosimilitud consiste en hallar un vector de parámetros  $\theta_{MV}$  (el *estimador de máxima verosimilitud* -EMV-) de manera tal que la suma

$$S^2(\theta) = \sum_t \frac{(\log(I_t) - \log(q) - \log(\overline{B}_t))^2}{\sigma_t^2}$$

sea mínima cuando  $\theta = \theta_{MV}$ , con el estimador de máxima verosimilitud de  $\log(q)$ :

$$\log(\hat{q}) = \frac{\sum_t \frac{1}{\sigma_t^2} \log\left(\frac{I_t}{\overline{B}_t}\right)}{\sum_t \frac{1}{\sigma_t^2}}$$

Una vez obtenidos los estimadores de máxima verosimilitud, y estimada la matriz de covarianza de dichos estimadores, la próxima etapa será la simulación de valores de biomasa, biomasa desovante, reclutamientos y de otros indicadores biológicos que interesen. Este procedimiento se llevará a cabo mediante simulación de Monte Carlo.

## Simulación de Montecarlo para la generación de vectores de parámetros

En este caso, pueden considerarse dos maneras de realizar una simulación de Monte Carlo: una manera es generar valores del índice y de los parámetros biológicos de entrada y luego estimar el vector de parámetros  $\theta$  para cada secuencia generada. La dificultad de este procedimiento es que para cada secuencia que se genere se deberá realizar la estimación con algún algoritmo de optimización numérica lo cual puede ser ineficiente cuando el número de parámetros es grande o el modelo considerado involucra la resolución de una gran cantidad de ecuaciones no lineales. Otra manera de realizar la simulación es estimando la matriz de covarianza asintótica de los estimadores de máxima verosimilitud  $\Omega$ . Como la distribución asintótica de los estimadores es normal, se pueden generar réplicas de los parámetros con distribución normal con media los valores estimados, y con varianza la matriz de covarianza asintótica estimada de dichos estimadores. Este último método puede ser más eficiente numéricamente respecto al anterior (exceptuando el problema de estimación de la matriz de covarianza discutido en [Aubone, 2010]) y es el que se utilizó en este trabajo.

De esta forma, la simulación de Monte Carlo permite generar réplicas del vector de parámetros. Estas réplicas son utilizadas luego para proyectar y calcular diferentes puntos biológicos de referencia y riesgos biológicos asociados.

## Estimación de la matriz asintótica de covarianza $\Omega$ de los EMV

Se tiene que  $\Omega \in \mathbb{R}^{n \times n}$ , donde  $n$  es el número total de parámetros del modelo. La estimación de la matriz de covarianza de los EMV puede ser un proceso complejo, sobre todo si la cantidad de parámetros involucrada es grande. Como el cálculo de esta matriz involucra estimaciones de derivadas parciales de segundo orden y operaciones entre matrices (incluida la inversión), a medida que aumenta el número de parámetros del modelo el problema crece. Si la matriz es aproximadamente singular, i.e.  $|\det(\Omega)|$  es pequeño, la inversión numérica de la misma puede fallar y esto es probable que ocurra si el modelo está sobre-parametrizado o si los valores estimados no son los de máxima verosimilitud, con lo que la estimación de esta matriz es un problema de cálculo numérico complejo.

Una vez obtenidos los estimadores de máxima verosimilitud, puede estimarse la matriz de covarianza asintótica de dichos estimadores  $\hat{\Omega} = (I_{\theta_{MV}})^{-1}$  con  $I_{\theta_{MV}}$  la matriz de información de Fischer calculada con los parámetros de máxima verosimilitud, de la siguiente manera:

- i) Se calcula primero la matriz hessiana  $H$  del logaritmo de la verosimilitud.
- ii) Se factoriza  $H$  por Cholesky, de manera que  $H = SS^T$  con  $S$  matriz triangular inferior, y  $S^T$  la matriz traspuesta de  $S$ . Si la matriz no fuera definida positiva se calcula la descomposición de Cholesky de la matriz  $H + \mu Id$  para valores de  $\mu$  pequeños [J. E. Dennis, 1987].
- iii) Se calcula la aproximación

$$\hat{\Omega} = (S^T)^{-1} S^{-1}$$

## Generación de réplicas del vector de parámetros

Fijado  $NSIM$  el número máximo de simulaciones a realizar (que es el número total de réplicas que se desean) y dados el vector de parámetros del modelo estimado por máxima verosimilitud  $\theta_{MV}$  y la matriz de covarianza asintótica de dichos estimadores  $\hat{\Omega}$ , la generación de una réplica del vector de parámetros se obtiene efectuando los siguientes pasos:

- i) Se generan vectores aleatorios  $u_j \sim \mathcal{N}(0, Id)$  con  $j = 1, \dots, NSIM$
- ii) Se consideran los vectores  $\theta_j = \hat{\theta} + S^{-1}u_j$ . Luego, se obtiene que  $\theta_j \sim \mathcal{N}(\hat{\theta}, \hat{\Omega})$
- iii) Se utilizan los valores de  $\theta_j$  para calcular todos los valores restantes deseados como por ejemplo: tasas instantáneas de mortalidad por pesca, tasas instantáneas de mortalidad natural, biomasa desovante, etc.

$C_{(t)}$	Vectores de captura por edad
$M_{(t)}$	Vectores de mortalidad natural
$PM_{(t)}$	Vectores de pesos medios poblacional
$PHM_{(t)}$	Proporciones de hembras maduras
$I_{(t)}$	Indice de abundancia
$F_0$	Vector de parámetros iniciales

CUADRO 1. Datos requeridos para ejecutar el software

Este procedimiento permite, por ejemplo, calcular intervalos de confianza para el vector de biomasa y biomasa desovante o para el vector de reclutamientos. A su vez, permite obtener réplicas del vector de tasas instantáneas de mortalidad por pesca del último tiempo en conjunto con réplicas apareadas del vector de estado poblacional en dicho tiempo del período de diagnóstico que es justamente el punto de partida para las proyecciones y el análisis de riesgo.

### Funcionamiento del módulo DiagSim

El módulo DiagSim es uno de los dos módulos que componen el FISHGURU+ que constituye una solución informática para realizar diagnóstico y simulación dado un modelo específico, en particular, en este caso, es el modelo explicado mas arriba. Los datos necesarios son los que pueden verse en el cuadro 1.

Dichos datos deben hallarse distribuidos en distintas hojas de una planilla de Microsoft Excel de la siguiente forma:

Hoja 1 - **Datos**: La primer hoja contiene datos dimensionales de las tablas a ingresar y número de simulaciones que ejecutará el programa. En la figura 1, se aprecia un ejemplo. En la celda B1 se ingresa el número de edades; en la celda B2 se ingresa el número de años; en la celda B3 se ingresa el número de simulaciones y por último en la celda B4 se ingresa el valor de  $\mu$  para aplicar el algoritmo de Cholesky modificado cuando ello fuese necesario.

	A	B	C
1	edades	7	
2	años	21	
3	NSIM	100	
4	mu	0.0003000000000000	
5			
6			
7			
8			
9			

FIGURA 1. Ejemplo de la hoja 1

Hoja 2 - **Capturas (en miles)**: La segunda hoja debe contener la matriz de captura por edades en miles de individuos, que debe tener a los datos de cada año de evaluación como columna y los datos de cada edad como fila. Los datos se ingresan a partir de la celda B3 como se ve en el ejemplo de la figura 2.

	A	B	C	D
1	CAPTURA POR EDAD en número de individuos			
2		1989	1990	1991
3	1	14927	328386	3276
4	2	189642	414599	4246
5	3	152881	322794	2938
6	4	99775	129621	1151
7	5	50422	42501	428
8	6	60001	16121	171
9	7	46880	11575	130
10				
11				

FIGURA 2. Ejemplo de la hoja 2

Hoja 3 - **Mortalidad natural**: La tercer hoja debe contener la matriz de las mismas dimensiones que la matriz de captura, que contenga los valores de mortalidad natural debiendo tener a los datos de cada año de evaluación como columna y los datos de cada edad como fila. Los datos se ingresan a partir de la celda B3 como se ve en el ejemplo de la figura 3.

	A	B	C	D
1	TASA INSTANTANEA DE MORTALIDAD			
2		1989	1990	1991
3	1	0.3	0.3	0.3
4	2	0.3	0.3	0.3
5	3	0.3	0.3	0.3
6	4	0.3	0.3	0.3
7	5	0.3	0.3	0.3
8	6	0.3	0.3	0.3
9	7	0.3	0.3	0.3
10				

FIGURA 3. Ejemplo de la hoja 3

Hoja 4 - **Pesos medios poblacionales al inicio del año**: La cuarta hoja debe contener una matriz de las mismas dimensiones que la matriz de capturas y debe tener por columnas a los vectores de pesos medios por edad que son válidos al principio del año correspondiente.

	A	B	C	D
1	PESOS MEDIOS POBLACIONALES AL			
2		1989	1990	1991
3	1	0.03490	0.03490	0.034
4	2	0.14500	0.14500	0.145
5	3	0.31460	0.31460	0.314
6	4	0.54320	0.54320	0.543
7	5	0.82750	0.82750	0.827
8	6	1.16470	1.16470	1.164
9	7	1.55210	1.55210	1.552
10				

FIGURA 4. Ejemplo de la hoja 4

Hoja 5 - **Pesos medios poblacionales a la mitad de año**: La quinta hoja debe ser igual a la cuarta pero con los datos de pesos medios por edad.

O38 <span>fx</span>				
	A	B	C	D
1	PESOS MEDIOS POBLACIONALES a M			
2		1989	1990	199
3	1	0.07976	0.07976	0.079
4	2	0.21523	0.21523	0.215
5	3	0.41387	0.41387	0.413
6	4	0.67447	0.67447	0.674
7	5	0.99614	0.99614	0.996
8	6	1.37818	1.37818	1.378
9	7	1.82003	1.82003	1.820
10				

FIGURA 5. Ejemplo de la hoja 5

Hoja 6 - **Proporción de hembras maduras**: La sexta hoja debe ser igual a la cuarta pero con los datos de proporción de hembras maduras por año.

A2 <span>fx</span>				
	A	B	C	D
1	PROPORCION DE HEMBRAS MADURA			
2		1989	1990	199
3	1	0.00000	0.00000	0.000
4	2	0.06576	0.06576	0.065
5	3	0.78981	0.78981	0.789
6	4	0.99500	0.99500	0.995
7	5	1.00000	1.00000	1.000
8	6	1.00000	1.00000	1.000
9	7	1.00000	1.00000	1.000
10				

FIGURA 6. Ejemplo de la hoja 6

Hoja 7 - **Indice**: La séptima hoja debe contar con los datos del índice medio en la tercer fila y de la dispersión del mismo en la cuarta fila. En aquellos años que no se cuente con valores de tal índice debe dejarse la celda vacía.

D20 <span>fx</span>				
	A	B	C	D
1	INDICES DE AB			
2	edad/añ	1989	1990	19
3	I	631.2532000	627.8783000	593.86
4	si	335.1895000	341.3124000	361.16
5				
6				
7				
8				
9				
10				
11				

FIGURA 7. Ejemplo de la hoja 7

Hoja 8 - **Vector inicial ( $F_0$ )**: La octava hoja debe tener un valor para el vector inicial ( $F_0$ ) que debe ubicarse en forma de fila a partir de la celda B2 como puede observarse en la figura 8



	H3		$f_x$	
	A	B	C	D
1	VECTOR INICIAL			
2	F0	0.2	0.17	0.98
3				
4				
5				
6				
7				
8				
9				
10				
11				

FIGURA 8. Ejemplo de la hoja 8

Las hojas deben encontrarse en este orden en el archivo de Microsoft Excel (como puede apreciarse en la figura 9)

Una vez ejecutado el software guarda los resultados en el subdirectorío donde se encuentra el archivo de entrada (la planilla de Microsoft Excel). Los archivos de salida son los que figuran en las tablas 2 y 3

21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				

Datos / C / M / PM0 / PM05 / PHM / I / F0

Listo

FIGURA 9. Orden de las hojas del archivo de entrada

Bs.jpg	Gráfico de la tendencia de biomasa
Fs.jpg	Gráfico de variación de la F del último tiempo
Is.jpg	Gráfico del ajuste del índice
Recs.jpg	Gráfico de la tendencia del reclutamiento
SBs.jpg	Gráfico de la tendencia de la biomasa desovante

CUADRO 2. Tabla de archivos de gráficos

Los archivos de datos tienen un formato particular de manera que puedan ser integrados de manera sencilla a otro software. En particular:

- SR.csv tiene en la segunda fila los valores (separados por comas) de un parámetro denominado bit y el número de simulaciones. El parámetro bit debe ser 1 si la función de reclutamiento es la de Beverton y Holt, 2 si es la función de Ricker y 3 si la función es constante.
- SBV.csv tiene en la segunda fila los valores (separados por comas) de un parámetro denominado bit y el número de simulaciones. El parámetro bit debe ser 1 si existe el dato de la biomasa desovante vírgen y 0 en caso contrario.



DatosSB.csv	Datos de la biomasa desovante
Fs.csv	Simulaciones del vector F del último tiempo
Ms.csv	Simulaciones (si correspondiese) del vector M del último tiempo
Ntm.csv	Simulaciones del vector Ntm del último tiempo
SVB.csv	Datos de la biomasa virgen
SR.csv	Simulaciones de los parámetros de reclutamiento

CUADRO 3. Tabla de archivos de datos

## REFERENCIAS

- [Aubone, 2010] Aubone, A. (2010). *Modelos Discretos de Dinámica de Poblaciones de Peces Explotadas*. Lecturas en biomatemática. Editorial Martin.
- [J. E. Dennis, 1987] J. E. Dennis, R. B. S. (1987). *Numerical methods for unconstrained optimization and non-linear equations*. Classics in applied mathematics 16. Society for Industrial and Applied Mathematics.
- [Pope, 1972] Pope, J. G. (1972). Investigation of the accuracy of virtual population analysis using cohort analysis. *International Commission For The Northwest Atlantic Research Bulletin*.