INTERACCIÓN ENTRE EXCEL Y EL LENGUAJE R, UNA APLICACIÓN EN BIOLOGÍA PESQUERA

INTERACTION BETWEEN EXCEL AND LANGUAGE R, AN APPLICATION IN FISHERY BIOLOGY

Raúl Carvajal[1,2], Juan Madrid[2], Pedro Valdez[2], Aníbal Zaldívar-Colado[1]
[1]Facultad de Informática Mazatlán, Universidad Autónoma de Sinaloa, Mazatlan, Sinaloa, México
[2]Centro Regional de Investigaciones Pesqueras, Instituto Nacional de Pesca, Mazatlán, Sinaloa, México
Email: rcarvajal@gmail.com

Resumen
En este trabajo se presenta la metodología básica para utilizar el paquete matemático – estadístico R y su lenguaje de programación, desde una hoja de cálculo de Excel. Se presenta un ejemplo en el cual, a partir de los datos de longitud provenientes de un muestreo de la captura de una determinada pesquería, se obtienen los estadísticos descriptivos y el histograma correspondiente. Se calcula la distribución de tallas en tiempos posteriores, utilizando la ecuación de crecimiento de Von Bertalanffy y la curva de mortalidad natural. Finalmente utilizando la relación peso – longitud se hace un estimado de la biomasa.

Palabras Clave: *Hoja de Cálculo, Excel, Lenguaje R, Dinámica de Poblaciones, Pesquerías*.

Abstract
In this paper we present the basic methodology to use the mathematical - statistical package R and its programming language, from an Excel spreadsheet. An example is presented in which, from the length data from a sampling of the catch of a certain fishery, the descriptive statistics and the corresponding histogram are obtained. The size distribution is calculated in later times, using the Von Bertalanffy growth equation and the natural mortality curve. Finally, using the weight - length relationship, an estimate of the biomass is made.

Keywords: *Spreadsheet, Excel, R Language, Population Dynamics, Fisheries*.

1 INTRODUCCIÓN

El uso de las hojas de cálculo, y particularmente de Excel, es muy popular para muy diversas aplicaciones en todos los ámbitos. A través de esta herramienta se pueden hacer gran variedad de cálculos y gráficas que resultan de gran utilidad. Asimismo el procesamiento estadístico de los datos reviste especial relevancia en ciertas áreas y resulta muy conveniente utilizar paquetes estadísticos diseñados para ello.

Entre los muchos paquetes estadísticos que se pueden conseguir en el mercado se encuentra TIBCO Spotfire S+®, que cuenta con su propio lenguaje de programación S y que es bastante completo [1]. Fue desarrollado por John Chambers de los Laboratorios Bell, hacia 1981 y se actualiza constantemente, incluyendo muchos módulos [2]. Sin embargo el costo de adquirir este producto puede ser oneroso para muchos.

Existe una alternativa compatible con S, que bajo la categoría de Software libre puede ser bajado gratuitamente de Internet [3]: El lenguaje R. Desarrollado por Robert Gentleman y Ross Ihaka de la Universidad de Auckland, Nueva Zelandia, fue puesto a disposición del público en 1995, como un proyecto GNU (Licencia Pública General, de la Fundación de Software Libre) [4].

R es un lenguaje de programación e interactivo para estadística y gráficas, similar al S. Provee una amplia variedad de técnicas estadísticas y gráficas (modelado lineal y no lineal, pruebas estadísticas clásicas, pruebas no paramétricas, análisis de series de tiempo, análisis multivariado: clasificación, clustering, etc.). Una de las fortalezas de R es la facilidad para producir gráficas con calidad para publicaciones bien diseñadas, incluyendo símbolos matemáticos y fórmulas.

En el Instituto Nacional de Pesca (Inapesca), se realiza un muestreo continuo de las diversas pesquerías de nuestro país, con el fin de evaluarlas. El programa de preferencia que es utilizado para capturar los datos, obtener los estadísticos descriptivos básicos y graficar, es sin duda alguna Excel. Son muy pocos los que llegan a utilizar un paquete estadístico, ya sea por no conocerlo, no tener la licencia para el uso del mismo, etc. Si bien la

mayoría de los paquetes estadísticos pueden importar/exportar datos desde/hacia Excel, no se ha generalizado su uso y se sigue trabajando en la tradicional hoja de cálculo, aunque se tengan que hacer manualmente diversos procedimientos que son rutinarios.

El proyecto R, tiene una enorme ventaja ante esta situación. Tiene un programa llamado RExcel [5], cuya presentación y forma de uso es idéntica a la del Excel común, pero que tiene unas funciones agregadas que le permiten conectarse con el lenguaje R, lo que le da un enorme valor agregado.

RExcel es una aplicación que permite ejecutar programas escritos en R desde Excel. Las principales características de RExcel son:

- Transferencia de datos entre R y Excel.
- Ejecutar código de R en Excel.
- Escribir macros llamando a Excel mediante funcionalidades de R sin que el usuario tenga que utilizar R.
- Llamar a funciones de R desde las celdas de Excel.
- Se instala mediante el paquete RExcelInstaller de la *Comprehensive R Archive Network* (CRAN).

Aprovechando estas características, en este trabajo se presenta un ejemplo de la interacción entre ambos programas. Se trata del análisis de los datos de longitud total correspondientes a los muestreos que se realizan con regularidad por el Inapesca. Además de calcular los estadísticos básicos, el histograma inicial y el diagrama de caja, se calcula el crecimiento en longitud utilizando la ecuación de Von Bertalanffy, la sobrevivencia y se obtiene la distribución de tallas para tiempos sucesivos y se hace una estimación de la biomasa disponible.

2 METODOLOGÍA

2.1 Instalación de R y RExcel

Antes que nada, hay que instalar el paquete y lenguaje de programación R en la computadora. Se puede instalar bajo plataformas Unix, Linux, MacOs y Windows, para ello basta visitar la página de Internet [3] y bajar el programa de la CRAN o de alguno de sus espejos (mirrors).

El RExcel funciona en Windows (XP o Vista) con Excel 2000, 2002, 2003 y 2007. Se integra completamente con el paquete Rcmdr de R (conocido como R commander) y se puede bajar accediendo a la página de Internet [6]. La forma más sencilla de instalarlo es bajando el paquete: RandFriendsSetup2100V3.0-18-1. Este programa instala todos los programas necesarios para el uso de RExcel.

En Linux sólo se puede usar con OpenOffice, con su hoja de cálculo Calc, mediante RCalc. También se puede instalar MSOffice bajo Wine en Linux, un cuasi emulador de Windows. Para MacOs existe un programa muy bueno que también funciona en Windows, que puede obtenerse en la página de Internet [7].

2.2 Transmisión de datos entre Excel y R

Una vez instalado el R y el RExcel, se puede usar el RExcel de manera idéntica que el Excel, pero ahora tendremos la posibilidad de enviar datos a R, procesarlos en R directamente o a través de algún programa, recibir gráficas hechas con R y pasar resultados obtenidos con R a nuestra hoja de cálculo. Esta transmisión de datos se puede realizar de dos formas: 1) Utilizando el menú extra llamado RExcel, opciones: *Get R Value* (para traer valores de R a Excel), *Put R Var* (pasa de Excel a R) o bien las instrucciones *#!rget* y *#!rput* que se ponen directamente en celdas de Excel y se ejecutan utilizando el menú extra con la opción *Run Code*. En ambos casos hay que dar el rango de celdas donde están los datos a mandar a R o donde se pondrán los provenientes de R y el nombre de variable con que lo identifica R. Por ejemplo para pasar a R los datos que están el rango de celdas de A10:C30 se selecciona ese rango, se va al menú extra RExcel opción *Put R Var* y se le da el nombre de la variable (en esto caso sería una matriz de 20 renglones y 3 columnas) con la cual será identificada en R, digamos *datos;* o bien se pone en tres celdas horizontales consecutivas lo siguiente:

#!rput datos =sheetrangeaddress(A10:C30)

Cuando se desee realizar la transferencia se seleccionan estas tres celdas se va al menú extra RExcel, opción *Run Code.*

Para ejecutar un programa en R previamente grabado, se pone el nombre del programa (que debe estar en el directorio de trabajo de R, o bien poner la trayectoria correspondiente) en una celda cualquiera; se selecciona y del menú extra RExcel, se elige la opción *Run* Code. Una referencia completa sobre el uso de RExcel y R, puede encontrarse en [5].

2.3 Aplicación Biológico – Pesquera

Para procesar los datos de las longitudes obtenidas del muestreo, se prediseñó una hoja de cálculo en RExcel, donde hay que poner los parámetros de las ecuaciones que se utilizarán para hacer los cálculos. Los datos que será necesario proporcionar son: Longitud infinito (L_{oo}), k y t_0 para la ecuación de crecimiento de Von Bertalanffy, la tasa instantánea de mortalidad natural, los valores de coeficiente a y exponente b de la relación peso – longitud, el número de unidades de tiempo a calcular y el ancho del intervalo de clase (para el histograma).

Los datos de longitud se capturan a partir de la celda A13 hacia abajo. Cuando se termina la captura se seleccionan las celdas que contienen estos datos y se envían a R, a través del menú extra RExcel, opción *Put R Var* dándole el nombre de *lt*. Los datos de los parámetros

de las ecuaciones se pasan seleccionando las celdas I16:K23, menú extra RExcel opción *Run Code*. En la celda C24 está en nombre del programa en R, se selecciona y otra vez *Run Code* y se desplegará el histograma, diagrama de caja, curva de crecimiento y distribución de tallas a diferentes tiempos. Para ver los resultados numéricos ir a la celdas H25:J27, seleccionar y *Run Code*. Los resultados numéricos se desplegarán en la hoja de cálculo.

Para los cálculos de los estadísticos básicos, histograma y diagrama de caja se utilizaron las funciones que provee el lenguaje R [8]

Para obtener crecimiento, mortalidad y distribución de tallas se aplicaron las siguientes ecuaciones [9]:

Crecimiento de Von Bertalanffy:

$$L_t = L_{oo}(1 - e^{-k(t-to)} \quad (1)$$

donde L_t es la longitud de un individuo a la edad t, t_o es el factor de corrección en el tiempo para el tamaño al nacimiento o reclutamiento, L_{oo} es la longitud asintótica teórica de un individuo (longitud máxima), y k es la constante de crecimiento.

Dado que en la ecuación (1) el tiempo debe expresarse como la edad del individuo y puesto que ese dato no está disponible, sino solamente la longitud total, se utilizará el criterio sugerido en [10] y es mostrado en la ecuación (2).

$$L_{t+1} = L_t + \Delta l \quad (2)$$

$$\Delta l = (L_{oo} + l*)(1 - e^{-k}) \quad (3)$$

es decir, la longitud al tiempo siguiente (t+1) es igual a la longitud actual más el cambio en longitud (Δl) para ese periodo, el cual se calcula utilizando la ecuación (3) donde l_* es la punto medio de cada intervalo de clase (marca de clase). La longitud al tiempo 1 corresponde al muestreo realizado.

La mortalidad natural se calcula de acuerdo a [11] y se muestra en la ecuación (4).

$$N(t) = N_0 e^{-zt} \quad (4)$$

donde *N(t)* es el número de individuos al tiempo t, N_0 es la cantidad de individuos al tiempo 0 y z es la tasa de mortalidad natural instantánea.

Una vez obtenido el crecimiento y mortalidad en un tiempo determinado, se obtiene la estadística básica se reagrupan las tallas y se hacen las gráficas correspondientes.

Posteriormente se utiliza la relación peso – longitud:

$$W = aL^b \quad (5)$$

donde a y b son parámetros correspondientes a la especie que se trate.

Finalmente se calcula la biomasa:

$$B(t) = N(t) * W(t) \quad (6)$$

donde *B(t)* es la biomasa en un tiempo dado, *N(t)* el número de individuos que se obtiene de la ecuación (4) y *W(t)* el peso promedio de los individuos para ese intervalo de clase, obtenido de la ecuación (5).

3 RESULTADOS

3.1 Programa en Lenguaje R

El lenguaje R es un intérprete y permite, como cualquier lenguaje de programación de uso general, operaciones aritméticas y lógicas, condiciones, bucles, definición de funciones, manejo de archivos, etc. Es además un lenguaje que maneja objetos (una matriz es un objeto, una gráfica, un *data frame*, etc.) y como está orientado al análisis matemático estadístico, tiene instrucciones simples y potentes que permiten hacer diversos tipos de gráficas y análisis estadísticos [12].

En la Figura 1 se muestra parte del código generado y puede apreciarse que se utilizan instrucciones de control de bucles (*for, while*), condiciones (*if*), asignaciones (<-), funciones estadísticas (*max, min, mean, var, median)* operaciones aritméticas y en los últimos renglones se realiza el histograma y el diagrama de caja, utilizando los métodos *hist* y *boxplot*, los cuales son bastante sencillos y potentes.

3.2 Resultados del programa R

En la Figuras 2 se presentan las gráficas desplegadas en ventanas generadas por R y en las Tablas I y II, se presentan los resultados que se depositan en la hoja de cálculo Excel a partir de cálculos realizados en R con el programa desarrollado.

En la Figura 2 puede observarse que, para los datos procesados, el histograma presenta una distribución aproximadamente normal, con algunos datos atípicos en el extremo derecho. Estos "datos extremos" (*outliers*) son detectados claramente con el diagrama de caja, se trata de 3 casos cuyos valores están arriba de 200 mm en este caso.

Los datos de la Tabla I presentan los estadísticos descriptivos básicos de la muestra, incluido el intervalo de confianza para la media poblacional μ.

La Tabla II presenta la distribución de frecuencias, tanto absoluta, como relativa (frecuencia absoluta /

tamaño de la muestra) y relativa acumulada para la muestra analizada.

La Figura 3 presenta en la parte de arriba las curvas de crecimiento correspondientes a la clase de longitud menor, la clase de longitud media y la clase de longitud mayor. Puede observarse que las tallas menores crecen más que las tallas mayores, de acuerdo a la ecuación de crecimiento, ya que en este caso las tallas mayores están cerca de la longitud máxima (L_{oo}).

```
#
## Calcula estadísticos básicos
#
maxl <- max(lt);  minl <- min(lt)
nl <- length(lt); proml <- mean(lt)
varl <- var(lt);  stdl <- sqrt(varl)
cvl <- stdl/proml*100; medianal <- median(lt)
errorl <- stdl/sqrt(nl)
prommin <- proml-1.96*errorl
prommax <- proml+1.96*errorl
#
## Datos para hacer tabla de frecuencias
#
lmin <- trunc(minl/ancho)*ancho
lmax <- trunc((maxl+ancho)/ancho)*ancho
if(lmin == minl) lmin <- lmin-ancho
ltord <- sort(lt);
lc <- seq(lmin,lmax,ancho); nc <- length(lc)
fab <- rep(0,nc);
#
## Hace tabla de frecuencias absolutas
#
j <- 1
for(i in 1:nl) {
        while(ltord[i] > lc[j]) j <- j+1
        fab[j] <- fab[j] + 1
}
#
## Genera matriz s para poner resultados en Excel
#
s <- matrix(lc,nc,1)
s <- cbind(s,fab,frel,facum)
#
## Hace histograma y diagrama de cajas
#
split.screen(c(2,1))
screen(1)
hist(lt, breaks=lc, col="blue", main="Histograma de los
datos", xlab="Longitud total(mm)", ylab= "Frecuencia
absoluta")
screen(2)
boxplot(lt, horizontal=TRUE)
```

Figura 2 Parte del código generado en lenguaje R para la aplicación en Biología Pesquera

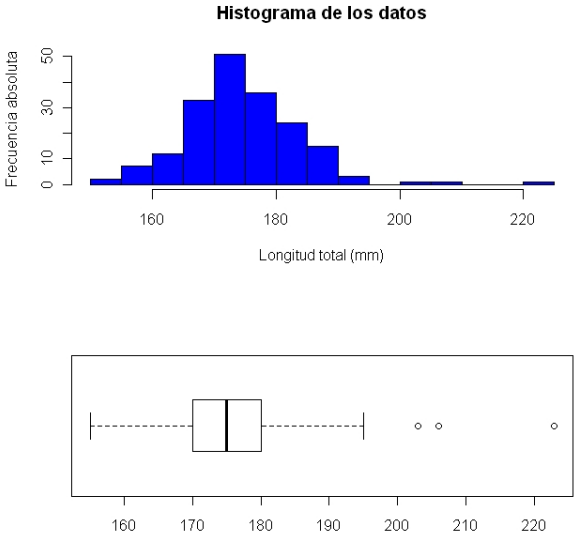

Figura 1 Histograma y diagrama de caja obtenido

Tabla 1 Estadísticos descriptivos que regresa el programa en la hoja Excel

n =	186
Promedio=	175.44
Varianza=	88.27
Desv. Std. =	9.40
Error Std. =	0.6889
C.V. =	5.36
Mediana =	175
Mínimo =	155
Máximo =	223
Rango =	68
LimInfProm	174.09
LimSupProm	176.79

Tabla 2 Distribución de Frecuencias de la muestra analizada

Long (mm)	Frec abs	Frec Rel	Frec acum
150	0	0.000	0.000
155	2	0.011	0.011
160	7	0.038	0.048
165	12	0.065	0.113
170	33	0.177	0.290
175	51	0.274	0.565
180	36	0.194	0.758
185	24	0.129	0.887
190	15	0.081	0.968
195	3	0.016	0.984
200	0	0.000	0.984
205	1	0.005	0.989
210	1	0.005	0.995
215	0	0.000	0.995
220	0	0.000	0.995
225	1	0.005	1.000

En la parte inferior de la Figura 3, se grafican las curvas de distribución de tallas para cada unidad de tiempo. Cada línea en color corresponde a la distribución de tallas esperada de acuerdo al crecimiento y mortalidad, en cada unidad de tiempo. En este caso se observan 8 curvas que fueron las unidades de tiempo que se registraron, o sea, 8 semanas.

En el caso de que la especie que se analiza sea camarón, también se calcula la distribución de tallas comerciales para cualquier tiempo *t*. La Tabla III y la Figura 4 muestran estos resultados.

La Tabla IV y la Figura 5 muestran los resultados de los cálculos de la biomasa relativa total, en este caso para una muestra de camarón con cabeza. También se muestra el valor en dinero que tendría esa biomasa en cada tiempo a precios de exportación.

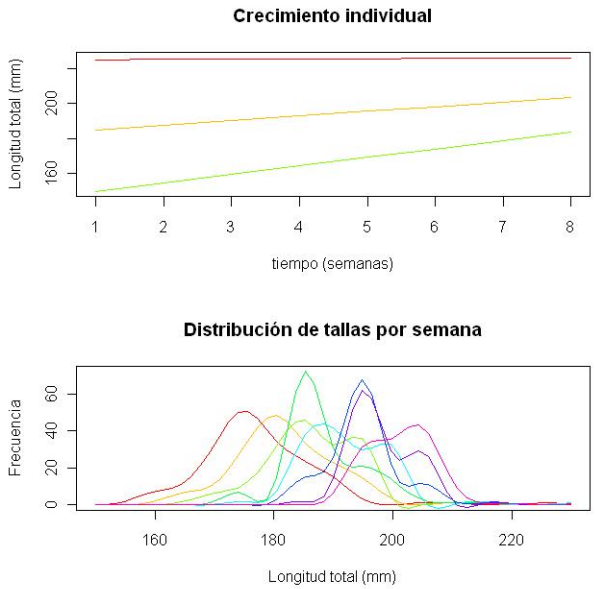

Figura 3 Curvas de crecimiento y distribución de tallas a diferentes tiempos

Tabla 3 Distribución de tallas comerciales de camarón

Talla	Porcentaje
U-10	10.61
U-12	13.64
U-15	30.30
16-20	27.27
21-25	15.15
26-30	1.52
31-40	0.00
41-50	1.52
51-60	0.00
61-70	0.00
71-80	0.00
80 Over	0.00

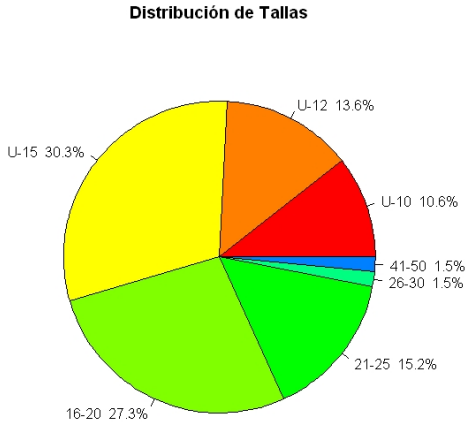

Figura 4 Diagrama de tallas comerciales de camarón

Tabla 4 Biomasa calculada y valor de la misma

tiempo (sem)	Biom Rel (%)	Valor relativo
1	100.0	100.00
2	102.1	109.86
3	102.5	120.00
4	99.0	119.06
5	98.0	122.15
6	96.1	128.80
7	94.3	127.71
8	92.3	123.44

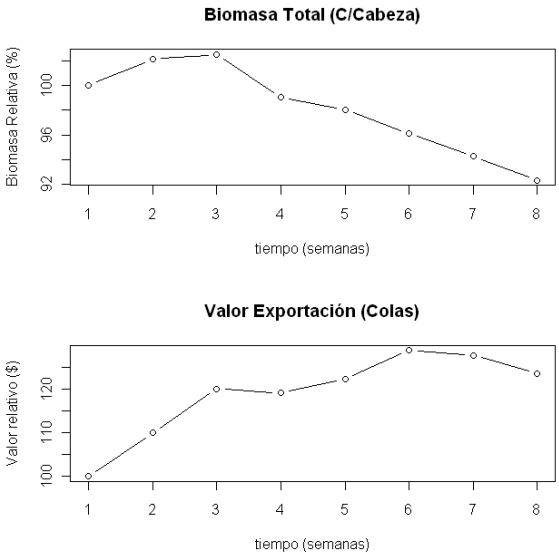

Figura 5 Gráfica de la Biomasa y su valor en el mercado internacional

4 DISCUSIÓN Y CONCLUSIONES

En este trabajo se ha presentado un ejemplo de un tipo de aplicación que se puede hacer desde la hoja de cálculo Excel, utilizando algunas de las herramientas del lenguaje R.

Como puede apreciarse resulta muy útil interactuar desde Excel con programas hechos en lenguaje R, que permiten la automatización de muchas tareas rutinarias y le permiten al usuario sentirse en un ambiente que conoce, sin tener necesidad de aprender un paquete estadístico y mucho menos programación.

Por supuesto que el desarrollo de los programas en R, lo deberá realizar algún usuario que conozca el lenguaje y la forma de transferir datos entre Excel y R, lo cual, como se ha descrito es sumamente fácil.

De esta forma, con sólo seleccionar celdas y elegir un menú se pueden realizar procesos que de otra forma requerirían de mayor tiempo y trabajo. Aún más, si se desea mejor presentación y facilidad para el usuario se pueden diseñar botones, menús y otras opciones de interfase gráfica con el usuario, a través de macros realizadas en *Visual Basic for Applications* (VBA), el estándar de macros para la familia de programas de *Office*.

La aplicación de Biología Pesquera aquí desarrollada, se realizaba comúnmente en Excel y el usuario podía tardarse horas en hacer todas las gráficas y cálculos. Con esta nueva aplicación en tan solo pocos minutos obtiene resultados. Y con la facilidad que mantiene sus datos en Excel y los resultados quedan también en la hoja de cálculo y por lo tanto los podrá usar para otras aplicaciones.

Finalmente podemos concluir que el uso del RExcel y la posibilidad de comunicarse con R para transferir datos entre las dos aplicaciones, así como el poder procesar los datos a través de programas desarrollados en lenguaje R, resulta de gran ventaja para los usuarios.

Con esta aplicación se pueden hacer simulaciones que permiten conocer como crecen los organismos, su talla comercial, la biomasa que representan y su valor en el mercado, lo que ayuda a tomar las mejores decisiones.

5 REFERENCIAS

[1] TIBCO Spotfire. URL:http://spotfire.tibco.com/Products/S-Plus-Overview.aspx. (31.11.2009).

[2] Becker, R. A.; Chambers, J. M.. A Language and System for Data Analysis. Bell Laboratories Computer Information Service, Murray Hill, New Jersey, 1981. [3] The R Project for Statistical Computing. URL:http://www.r-project.org. (31.11.2009).

[3] Ihaka, R; Gentleman, R. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5, 3(1996), pp. 299-314.

[4] Heiberger, R. M.; Neuwirth, E.; Through, R. Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics. Springer Verlag, New York, 2009.

[5] Baier, T.; Neuwirth, E. Statconn projects. URL:http://rcom.univie.ac.at/. (04.11.2009).

[6] The European Bioinformatics Institute. URL:http://www.ebi.ac.uk/microarray-srv/frontendapp/rworkbench.jnlp. (04.11.2009).

[7] Dalgaard, P. Introductory Statistics with R, 2nd ed. Springer, New York, 2008.

[8] Haddon, M. Modelling and quantitative methods in fisheries. Chapman & Hall/CRC, New York, 2001.

[9] Sullivan, P. J.; Lai, H. L.; Gallucci, V. F. A catch-at-length analysis that incorporates a stochastic model of growth.

[10] Can. J. Fish. Aquat. Sci. 47 (1990), pp. 184-198.

[11] Sparre, P.; Venema, S.C. Introducción a la evaluación de recursos pesqueros tropicales. FAO, Roma, 1997.

[12] Chambers, J. M. Software for Data Analysis: Programming with R. Springer, New York, 2008.

.