SIEVE BOOTSTRAP EN SERIES DE TIEMPO DE NUBOSIDAD EN EL CARIBE

Por

Walter Quispe Vargas

Tesis sometida en cumplimento parcial de los requisitos para el grado de

MAESTRO EN CIENCIAS
en
MATEMATICA
(Estadística)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ
Julio, 2006

Aprobada por:

_____	_____
Nazario Ramírez Beltrán, Ph.D. Presidente, Comité Graduado.	Fecha
_____	_____
Tokuji Saito, Ph.D. Miembro, Comité Graduado.	Fecha
_____	_____
Edgar Acuña Fernández, Ph.D. Miembro, Comité Graduado	Fecha
_____	_____
Nilda E. Aponte Avellanet, Ph.D. Representante de Estudios Graduados.	Fecha
_____	_____
Arturo Portnoy, Ph.D. Director Interino del Departamento.	Fecha

ABSTRACT

The Sieve Bootstrap is a resampling method, designed to deal with autocorrelated data, specifically a sequence of information taken at equal time intervals. Formally, the Sieve Bootstrap approximates a linear process by a sequence of autoregressive processes of order $p = p(n)$, where $p(n) \to \infty$, $\frac{p(n)}{n} \to 0$ as the sample size $n \to \infty$ for a time series that is expressed by an autoregressive $\mathrm{AR}\left(p(n)\right)$ model, it should be noted that the bootstrap is constructed over the residuals. In this thesis, we apply the Sieve Bootstrap to the construction of prediction intervals of cloudiness time series on Caribbean, which were obtained from the data base of level D2-DATA of the International Satellite Cloud Climatology Project. Results obtained show that the Sieve Bootstrap method provides a better prediction interval coverage. However the Box Jenkins technique shows a significance reduction in the length of the prediction intervals.

RESUMEN

El *Sieve Bootstrap* es un método de remuestreo diseñado especialmente para datos cuya cualidad es la dependencia entre sí, específicamente datos tomados a través del tiempo. Formalmente el *Sieve Bootstrap* aproxima un proceso lineal, mediante una sucesión de procesos autoregresivos de orden $p = p(n)$, donde $p(n) \to \infty$ y $\frac{p(n)}{n} \to 0$ cuando el tamaño de muestra $n \to \infty$, para una serie de tiempo, la cual se expresa mediante un modelo autorregresivo de orden $p(n)$, $\mathrm{AR}\left(p(n)\right)$. Es conveniente notar que el remuestreo se realiza en los residuales. En este trabajo se presenta el *Sieve Bootstrap*, para la construcción de intervalos de predicción de series de tiempo de nubosidad en el Caribe, usando la base de datos de niveles D2-DATA, extraídos del *International Satellite Cloud Climatology Project*. Los resultados obtenidos muestran que el método Sieve Bootstrap provee una mejor cobertura en los intervalos de predicción. Sin embargo la técnica Box Jenkins muestra una reducción significativa en la amplitud de los intervalos de predicción.

To my Wife and my Parents

AGRADECIMIENTOS

Al Dr. Nazario Ramírez por su gran ayuda, interés y sugerencias en el desarrollo del presente documento.

Al Dr. Andrés Alonso, del departamento de Estadística Universidad Carlos III de Madrid España, por su solidaridad, en el desarrollo del presente documento.

A mis Maestros del Recinto por las sabias enseñanzas que me impartieron.

A todos mis amigos y compañeros.

TABLA DE CONTENIDOS

LISTA DE TABLAS

Tablas **Página**

LISTA DE FIGURAS

Figuras **Página**

LISTA DE SIMBOLOS Y ABREVIATURAS

$\mathbb{N}$	Conjunto de Números Naturales.
$\mathbb{Z}$	Conjunto de Números Enteros.
$\phi(B)$	Polinomio Autorregresivo Regular.
$\Phi(B)$	Polinomio Autorregresivo Estacional.
$\theta(B)$	Polinomio de Media Móvil Regular.
$\Theta(B)$	Polinomio de Media Móvil Estacional.
$B(\cdot)$	Operador de Retardos.
$N(0,\sigma_\varepsilon^2)$	Distribución Normal con Media cero y Varianza σ_ε^2.
$RB(0,\sigma^2)$	Ruido Blanco con Media cero y Varianza σ^2.
T_0	Espacio Paramétrico.
Ω	Espacio Muestral.
F	Sigma Algebra.
$\gamma(\cdot)$	Función de Autocovarianza.
$\rho(\cdot)$	Función de Autocorrelación.
Δ^d	Operador de Diferencias Regulares de orden d
Δ_s^D	Operador de Diferencias Estacionales de orden D y periodo s.
λ	Parámetro de Transformación Box Cox.
$o(\cdot)$	Orden de Convergencia.
ISCCP	International Satellite Cloud Climatology Product.
AR(p)	Proceso Autorregresivo de orden p.
MA(q)	Proceso de Media Móvil de orden q.
ARMA(p,q)	Proceso Autorregresivo de Media Móvil de orden p,q.
ARIMA(p,d,q)	Proceso Autorregresivo Integrado de Media Móvil de orden p,d,q.
SARIMA(p,d,q)(P,D,Q)	Proceso Autorregresivo Integrado de Media Móvil Estacional de ordenes p,d,q,P,D,Q.
IID, iid	Independiente e Idénticamente Distribuidos.
AIC	Akaike Information Criterion.
AICC	Akaike Information Criterion Corrected.
BIC	Bayesian Information Criterion.

1 INTRODUCCION

El método de remuestreo *Bootstrap* propuesto por *Efron* (1979), es un procedimiento no paramétrico muy eficiente para estimar la distribución de una variable estadística. Sin embargo, al ignorar el orden de las observaciones, como ocurren en el caso de las observaciones dependientes, el bootstrap usualmente falla por que se rompe la estructura de dependencia. Es así que dentro de las series de tiempo estacionarias surge una aproximación basada en un modelo, el cual remuestrea residuos aproximadamente independientes e idénticamente distribuidos [Kreiss y Franke, 1992].

El método de *Sieve Bootstrap* propuesto por *Bühlmann* (1997) toma esta idea, ajusta un modelo paramétrico y luego remuestrea los residuales de una secuencia de aproximaciones de procesos autorregresivos para $\{X_t, t \in \mathbb{Z}\}$ con orden $p = p(n)$, el cual depende del tamaño de muestra *n*, puesto que al incrementar el tamaño de muestra *n*, el orden *p* se incrementa pero no con la misma rapidez como lo hace *n* [Bühlmann, 1997]. Esto es, en vez de considerar un modelo fijo finito dimensional, aproxima un modelo no paramétrico infinito, mediante una secuencia de modelos parametricos finito dimensionales.

En el análisis de series de tiempo es importante predecir valores futuros de una serie observada sobre la base de valores pasados, y más específicamente como calcular intervalos

de predicción; una aproximación es asumir que una serie de tiempo sigue un modelo lineal finito dimensional con la distribución de los errores conocidos, usualmente se asumen que es un proceso gausiano [*Box y Jenkins*, 1976]. En adición el *bootstrap* ha sido propuesto para usar distribuciones alternativas de los errores; de esta forma el *sieve bootstrap* se extiende para la construcción de intervalos de predicción para una clase de modelos lineales que incluye procesos Autoregresivos de Medias Móviles (ARMA) estacionarios e invertibles.

1.1 Motivación

En la construcción de modelos para series de tiempo de nubosidad se hace imprescindible contar con herramientas más sofisticadas, cuya tarea principal es encontrar predicciones o pronósticos, las cuales tengan un intervalo de confianza más pequeño comparado con los métodos tradicionales; para lograr este objetivo se comparan dos metodologías estadísticas que son: Los modelos de series de tiempo y el *Sieve Bootstrap*.

La estructura del *Sieve Bootstrap* por su naturaleza al remuestrear los errores, no asume ningún tipo de distribución conocida para estos, más aún utiliza la distribución empírica para estos, que después de un intenso remuestreo produce buenos resultados, teniendo en cuenta que dicha distribución no es conocida; de esta forma el *Sieve Bootstrap* en la construcción de intervalos de confianza para valores futuros de una serie de tiempo tiene un comportamiento no paramétrico.

1.2 Objetivos

- Estimar la distribución de las predicciones o valores futuros de las series de tiempo climatologicas que exhiben un comportamiento autorregresivo, y de esta manera construir intervalos de predicción que sea precisos y que tengan una amplitud mínima, a través del procedimiento *Sieve Bootstrap* en la construcción de intervalos de predicción.

- Construir un modelo adecuado para las series de tiempo de nubosidad a emplear, tanto para describir las series, para realizar predicciones de valores futuros y para la construcción de intervalos de predicción de las series utilizando la metodología *Sieve Bootstrap.*

- Implementar un algoritmo para calcular intervalos de predicción bajo la metodología de *Sieve Bootstrap*, y finalmente comparar la precisión de cada método, usando conjuntos de datos que caracteriza el comportamiento de las nubes en el caribe; los cuales son extraídos del *International Satellite Clouds Climatology Project* (ISCCP), y los datos correspondientes al producto D2-DATA.

1.3 Resumen de los Capítulos

La presente tesis esta estructurada en seis capítulos. El capítulo 2 proporciona algunos conceptos básicos relacionados a la teoría de series de tiempo. En el capítulo 3 se definen de manera formal y detallada la metodología *Bootstrap* para variables independientes e idénticamente distribuidas, así como la inadecuidad de éste para datos dependientes, y una alternativa para tratar este problema. En el capítulo 4 se presenta el *Sieve Bootstrap* como una solución para datos dependientes, su definición y una variante para tratar el problema de la construcción de intervalos de predicción de una serie temporal. El capítulo 5 muestra los resultados obtenidos usando el *Sieve Bootstrap* en intervalos de predicción usando conjuntos de datos del producto D2-DATA del *ISCCP*. Finalmente en el capítulo 6 se indican las conclusiones obtenidas y se mencionan algunos trabajos futuros de investigación sobre el *Sieve Bootstrap* para un modelo de función de transferencia.

2 ASPECTOS TEORICOS

2.1 Introducción

En los últimos años, el análisis de series de tiempo, se ha convertido en un método poderoso para analizar conjuntos de datos, cuya característica fundamental, es que se hayan observado a través del tiempo y a intervalos iguales, así los datos usualmente son dependientes entre sí, y cuyos principales objetivos es describir el comportamiento típico de una serie de tiempo $X_t = P + T + E$, donde P es el componente periódico, T es el componente de tendencia y E es el componente estocástico; y realizar predicciones o interpolaciones lo más confiable posibles, sin importar el área de investigación en que se encuentra.

Por otro lado las características ineludibles de una serie temporal, obligaron prácticamente a los académicos, a desarrollar nuevos y apropiados métodos para la inferencia estadística. La econometría fue la más privilegiada con el tratamiento de las series de tiempo, es así que en el año 2003, *Engle, R.F. y Granger C.* recibieron conjuntamente un "Premio Nobel" en economía cuyos avances fueron dados en econometría, por haber desarrollado métodos de análisis temporales con tendencias comunes *Cointegration and Methods of analyzing economic time series with time-varying volatility* (ARCH). En los setentas surge un gran texto, cuyos autores, *Box and Jenkins* (1976) establecen los principios fundamentales sobre el tratamiento de las series de tiempo.

En particular los modelos estacionales autorregresivos integrados de medias móviles *SARIMA(p,d,q)x(P,D,Q)$_S$,* son una gran generalización del estudio *Box-Jenkins* que se presenta brevemente. La metodología de *Box Jenkins* asume que una serie $\{X_t\}_{t\in Z}$, donde $\mathbb{Z}$ es el conjunto de los enteros, sigue un modelo lineal finito dimensional con la distribución de los errores conocido usualmente se asume un proceso gausiano. Así el modelo *SARIMA(p,d,q)(P,D,Q)$_S$* es: $\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = \theta(B)\Theta(B^s)\varepsilon_t$, donde: $\phi(B) = 1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p$ y $\Phi(B) = 1-\Phi_1 B^s-\Phi_2 B^{2s}-\ldots-\Phi_P B^{sP}$, son los operadores ó polinomios autorregresivos, asociados al componente regular y estacional respectivamente, además $\theta(B) = 1+\theta_1 B+\theta_2 B^2+\ldots+\theta_q B^q$ y $\Theta(B) = 1+\Theta_1 B^s+\Theta_2 B^{2s}+\ldots+\Theta_Q B^{sQ}$, son operadores ó polinomios de medias móviles, asociados al componente regular y estacional respectivamente. En adición *p, P, q, Q,* son los parámetros estructurales del modelo, donde *p* y *P* son los ordenes de los polinomios autorregresivos regular y estacional respectivamente, *q* y *Q* son los ordenes de los polinomios de medias móviles regular y estacional respectivamente. Los parámetros *d* y *D* representan el número de diferencias regular y estacional respectivamente, que son necesarias para lograr que la serie de tiempo se convierta en un proceso estacionario; y *s* es la magnitud del periodo. Por otro lado B es el operador de retrasos, definido por: $BX_t = X_{t-1}$ y $B^s X_t = X_{t-s}$, finalmente $\{\varepsilon_t\}_{t\in Z} iid \sim N(0,\sigma_\varepsilon^2)$, esto es que los errores del modelo se supone que son independiente e idénticamente distribuidos con una distribución normal con esperanza metafórica cero y varianza σ_ε^2 constante. Si *d, D* y *s* se conocen, típicamente el procedimiento de máxima verosimilitud, es empleado para estimar el resto de los parámetros del modelo. Si son desconocidos los parámetros se buscan

mediante la minimización de algunos criterios tales como: *Akaike* o *Bayesian Information Criterion.*

Los valores futuros o predicciones de la serie, se obtienen de la información histórica y con la estructura del modelo. Estas predicciones están dentro del intervalo de predicción: $\hat{E}[X_{T+h} \mid X_1 \ldots X_T] \pm Z_{\alpha/2}(\hat{\sigma}_\varepsilon^2 \sum_{j=0}^{h-1} \hat{\psi}_j^2)^{1/2}$, donde $\hat{E}[X_{T+h} \mid X_1 \ldots X_T]$ es el predictor lineal usando el modelo $\phi(B)\Phi(B)(1-B)^d(1-B^s)^D X_t = \theta(B)\Theta(B)\varepsilon_t$; $\hat{\psi}_j$ son los coeficientes estimados de la representación de medias móviles; $\hat{\sigma}_\varepsilon^2$ es la varianza estimada de los errores y $Z_{\alpha/2}$ es la $\alpha/2$-ésima cuantila de la distribución normal estándar.

2.2 Procesos Estocásticos

En el análisis de series de tiempo es necesario seleccionar un modelo matemático adecuado para la secuencia observada, que nos permita conocer la naturaleza de los valores futuros ó predicciones; de esta manera suponemos que cada observación x_t es un valor realizado de cierta variable aleatoria X_t, así una serie de tiempo $\{x_t, t \in T_0\}$ es entonces una realización de la familia de variables aleatorias $\{X_t, t \in T_0\}$, donde T_0 representa el espacio de parámetros.

Definición.- Un proceso estocástico es una familia de variables aleatorias $\{X_t, t \in T\}$ definido sobre un espacio de probabilidad (Ω, F, P), donde Ω es el espacio muestral, F es un sigma álgebra y P representa la probabilidad.

Como X_t es una variable aleatoria, para cada $t \in T$, X_t es una función $X_t(\cdot)$ sobre el conjunto Ω. Por otro lado, para cada $\omega \in \Omega$ fijo, $X(\omega)$ es una función sobre T.

Definición.- Las funciones $\{X(\omega), \omega \in \Omega\}$ sobre T, son conocidas como realizaciones de un proceso $\{X_t, t \in T\}$. En adelante usaremos el término de serie de tiempo como una realización de un proceso estocástico.

Definición.- Sea T el conjunto de todos los vectores $\{\mathbf{t} = (t_1, \ldots, t_n) \in T^n : t_1 < t_2 < \ldots < t_n, n = 1, 2, \ldots\}$. Entonces la función de distribución finita dimensional de $\{X_t, t \in T\}$ son las funciones $\{F_{\mathsf{t}}(\cdot), \mathbf{t} \in \mathsf{T}\}$, definidas por: $F_{\mathsf{t}}(\mathsf{x}) = P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n)$, donde $\mathsf{x} = (x_1, \ldots, x_n)' \in \mathrm{R}^n$, $\mathbf{t} = (t_1, \ldots, t_n)$.

2.3 Proceso Estacionario y Estrictamente Estacionario

La obtención de las distribuciones de probabilidad del proceso es posible en ciertas situaciones, sin embargo sólo podemos observar una realización del proceso estocástico que es la serie de tiempo. El proceso estocástico existe conceptualmente, pero no es posible obtener muestras sucesivas o realizaciones independientes del mismo, para poder estimar las características intrínsecas del proceso tales como su media, varianza, etc. A partir de su evolución es necesario suponer que las propiedades (distribución de las variables en cada instante) son estables a lo largo del tiempo, esto conduce al concepto de proceso estacionario.

Definición.- Si $\{X_t, t \in T\}$ es un proceso tal que $Var(X_t) < \infty$ para cada $t \in T$, entonces la Función de Autocovarianza $\gamma_X(\cdot,\cdot)$ de $\{X_t\}$ esta definido por $\gamma_X(r,s) = Cov(X_r, X_s) = E[(X_r - EX_r)(X_s - EX_s)]$, $r, s \in T$

Definición.- Un proceso estocástico $\{X_t, t \in T\}$ es estrictamente estacionario si la distribución conjunta de probabilidad, de $(X_{t_1},\ldots,X_{t_k})'$ y $(X_{t_1+h},\ldots,X_{t_k+h})'$son las mismas para todo entero positivo k y para todo $t_1,\ldots,t_k,h \in T$, esto es $F(X_{t_1},\ldots,X_{t_k})' = F(X_{t_1+h},\ldots,X_{t_k+h})'$.

Definición.- Un proceso estocástico $\{X_t, t \in T\}$, se dice que es estacionario ó estacionario de segundo orden si:

(i) $E|X_t|^2 < \infty$ para todo $t \in T$

(ii) $EX_t = \mu$ constante para todo $t \in T$

(iii) $\gamma_X(r,s) = \gamma_X(r+t, s+t)$ para todo $r, s, t \in T$

Definición.- El proceso $\{X_t, t \in T\}$ es Gausiano si y solamente si las funciones de distribución de $\{X_t, t \in T\}$ son todas normales multivariadas. Si un proceso gausiano es estacionario, entonces el proceso será estrictamente estacionario.

En la parte restante de este documento se usará la siguiente notación: el parámetro t (tiempo) discreto, esto es $t \in \mathrm{Z} = \{0, \pm 1, \pm 2, \ldots\}$ y el proceso será escrito como $\{X_t, t \in \mathrm{Z}\}$.

2.4 Propiedades de la función de autocovarianza

Sea $\{X_t, t \in \mathrm{Z}\}$ un proceso estacionario con función de autocovarianza $\gamma_X(r,s) = \gamma_X(r-s,0) = \gamma_X(h,0) = \gamma_X(h) = Cov(X_{t+h}, X_t)$, para todo $r,s,h,t \in \mathrm{Z}$, llamaremos a $\gamma_X(h)$ la función de autocovarianza con retardo h.

Proposición.- La función de autocovarianza $\gamma(\cdot)$ de un proceso estacionario $\{X_t, t \in \mathrm{Z}\}$ cumple las siguientes propiedades:

(i) $\gamma(0) \geq 0$.

(ii) $|\gamma(h)| \leq \gamma(0)$, para todo $h \in \mathrm{Z}$.

(iii) $\gamma(h) = \gamma(-h)$, para todo $h \in \mathrm{Z}$.

Prueba.

La primera propiedad es un hecho obvio, ya que por definición $Var(X_t) \geq 0$. En la segunda propiedad utilizaremos la desigualdad de *Cauchy-Schwarz*.

$|\gamma(h)| = |Cov(X_{t+h}, X_t)| \leq (Var(X_{t+h}))^{1/2} (Var(X_t))^{1/2} = \gamma(0)$

La tercera propiedad establecida como. $\gamma(-h) = Cov(X_{t-h}, X_t) = Cov(X_t, X_{t+h}) = \gamma(h)$ □

Observación: (Ergodicidad) Típicamente una función de autocovarianza de un proceso estacionario tiende a cero esto es: $\lim_{h\to\infty} \gamma(h) = \lim_{h\to\infty} \mathrm{cov}(X_{t+h}, X_t) = 0$.

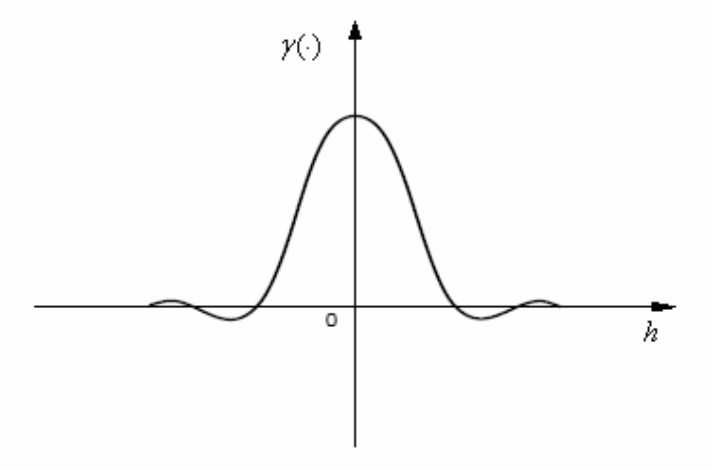


Figura 2.1: Función de Autocovarianza $\gamma(\cdot)$

Definición.- El proceso $\{X_t, t \in T\}$ con función de autocovarianza $\gamma_X(h)$, tiene la función de autocorrelación definida como: $Corr(X_{t+h}, X_t) = \rho_X(h) = \gamma_X(h)/\gamma_X(0)$, con $|\rho_X(h)| \leq 1$, para todo $t, h \in \mathrm{Z}$.

En adición a la autocorrelación entre X_{t+h} y X_t, es necesario investigar la correlación entre X_{t+h} y X_t después de que sus dependencias lineales sobre las variables que intervienen $X_{t+1}, X_{t+2}, \ldots, X_{t+h-1}$ haya sido removida. Así surge la correlación condicional: $Corr(X_{t+h}, X_t \mid X_{t+1}, \ldots, X_{t+h-1})$ denominada Autocorrelación parcial en el análisis de series de tiempo.

2.5 Función de autocovarianza muestral de una serie

En un proceso estacionario $\{X_t, t \in \mathrm{Z}\}$, frecuentemente observamos una serie de tiempo $\{X_1, X_2, \ldots, X_T\}$ de la cual estimamos la función de autocovarianza $\gamma(\cdot)$, consecuentemente la función de autocorrelación $\rho(\cdot)$, teniendo en cuenta la estructura de dependencia del proceso.

Definición.- La función de autocovarianza muestral de $\{X_1, X_2, \ldots, X_T\}$ esta definido por: $\hat{\gamma}(h) = (1/T)\sum_{j=1}^{T-h}(x_{j+h} - \bar{x})(x_j - \bar{x})$, y $\hat{\gamma}(h) = \hat{\gamma}(-h)$ para $0 \le h < T$, donde $\bar{x}$ es la media muestral definida por $\bar{x} = (1/T)\sum_{j=1}^{T} x_j$.

Una consecuencia importante es la función de autocorrelación muestral definida por: $\hat{\rho}(h) := \hat{\gamma}(h)/\hat{\gamma}(0), \quad |h| < T$.

2.6 Procesos Lineales Estacionarios

Una gran familia de procesos lineales estacionarios paramétricos son utilizados con gran frecuencia en el estudio de las series de tiempo, los más usados son los procesos autoregresivos (AR), medias móviles (MA), y la combinación de estos, llamados procesos ARMA, los que tienen una característica importante dentro de la teoría del predictor lineal.

2.6.1 Secuencia Aleatoria y Ruido Blanco

Consideremos $\{X_n, n = 1, 2, \ldots\}$ una secuencia de variables aleatorias definidas en el mismo espacio muestral Ω. Aquí $T = \{1, 2, \ldots\}$ y así tenemos un proceso de parámetro discreto o una secuencia aleatoria. Para todo $n \geq 1$ podemos escribir:

$$P\{X_1 = a_1, \ldots, X_n = a_n\} = P\{X_1 = a_1\} \times P\{X_2 = a_2 \mid X_1 = a_1\} \times \ldots \times P\{X_n = a_n \mid X_1 = a_1, \ldots, X_{n-1} = a_{n-1}\} \quad \textbf{(2.1)}$$

Donde los a_i representan el espacio de estados que pueden ser tomados como el conjunto de los reales. El caso más simple es cuando tenemos una secuencia $\{X_n, n = 1, 2, \ldots\}$ de variables aleatorias mutuamente independientes en el cual (2.1) se puede escribir como

$$P\{X_1 = a_1, \ldots, X_n = a_n\} = P\{X_1 = a_1\} \times P\{X_2 = a_2\} \times \ldots \times P\{X_n = a_n\} \quad \textbf{(2.2)}$$

Si las variables aleatorias $X_1, X_2, \ldots,$ tiene todas la misma distribución, entonces tenemos una secuencia de variables aleatorias independientes e idénticamente distribuidas (*iid*). En este caso el proceso X_n es estacionario, si $E(X_n) = \mu$, y $Var(X_n) = \sigma^2$ para todo $n \geq 1$, entonces

$$\gamma(h) = Cov(X_n, X_{n+h}) = \begin{cases} \sigma^2, & \text{si } h = 0 \\ 0, & \text{si } h \neq 0 \end{cases} \quad \textbf{(2.3)}$$

Definición.- El proceso $\{\varepsilon_t; t \in \mathbb{Z}\}$ se llama ruido blanco discreto si las variables aleatorias ε_t son no correlacionadas, esto es $Cov(\varepsilon_t, \varepsilon_s) = 0$ para $t \neq s$.

Tal proceso será estacionario si $E(\varepsilon_t) = \mu$, y $Var(\varepsilon_t) = \sigma^2$, para todo t, si su función de autocovarianza está dada por (2.3).

Obviamente si ε_t son variables aleatorias independientes, entonces también son no correlacionadas. Una secuencia de variables aleatorias *iid*, es llamada proceso puramente aleatorio.

De ahora en adelante representamos $\{\varepsilon_t; t \in \mathbb{Z}\}$ como un ruido blanco y suponemos $E(\varepsilon_t) = \mu = 0$, y se representa por $\{\varepsilon_t\} \sim RB(0, \sigma^2)$ y en el caso de un proceso puramente aleatorio se representa por $\{\varepsilon_t\} \sim i.i.d.(0, \sigma^2)$.

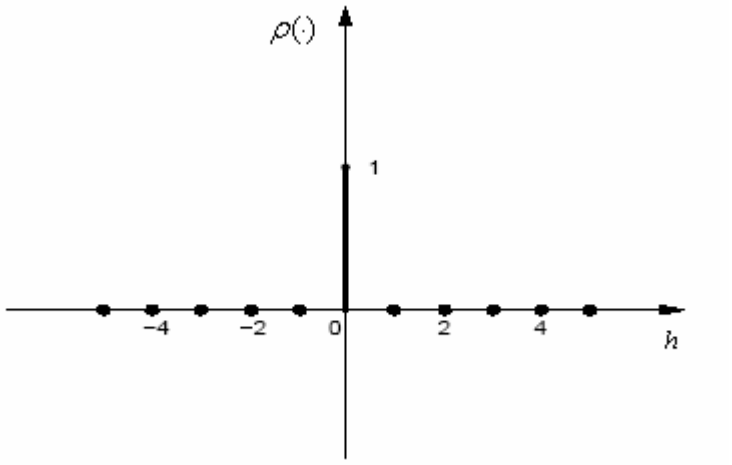


Figura 2.2: Función de Autocorrelación de un Ruido Blanco.

2.6.2 Procesos Autorregresivos AR(p)

Se dice que $\{X_t, t \in \mathrm{Z}\}$ es un proceso autoregresivo de orden *p*, se denota por $\{X_t\} \sim AR(p)$ y satisface la ecuación de diferencias:

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \ldots + \phi_p(X_{t-p} - \mu) + \varepsilon_t \qquad \textbf{(2.4)}$$

Donde $\mu, \phi_1, \phi_2, \ldots, \phi_p$ son parámetros del modelo y $\{\varepsilon_t\} \sim RB(0, \sigma^2)$. Si seguimos que $E(X_t) = \mu$, y escribimos el proceso de la forma:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t \qquad \textbf{(2.5)}$$

Entonces $E(X_t) = \mu = \dfrac{\phi_0}{1 - \phi_1 - \phi_2 - \ldots - \phi_p}$, y sin perder la generalidad es supuesta cero.

Definiendo el operador de retardo B, a través de $B^s X_t = X_{t-s}$ para $s \geq 1, s \in \mathbb{N}$, entonces si $\mu = 0$, (2.4) puede ser escrita como:

$$\phi(B) X_t = \varepsilon_t \qquad \textbf{(2.6)}$$

Donde $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ es el operador autoregresivo o polinomio autorregresivo de orden *p*, al proponer una solución para (2.6) del siguiente modo:

$\phi(B) X_t = \varepsilon_t \;\; \Rightarrow X_t = \phi(B)^{-1} \varepsilon_t \;\;\; \Rightarrow X_t = \psi(B) \varepsilon_t$.

Así la solución es:

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \qquad \textbf{(2.7)}$$

Donde $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots$, y debemos suponer que $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ para que (2.7) sea una solución estacionaria, ó el proceso *AR(p)*, definido por la ecuación $\phi(B) X_t = \varepsilon_t$ se dice ser causal; por otro lado el proceso $\{X_t\} \sim AR(p)$ que acepta la notación (2.7) se dice que tiene representación de media móvil infinita. Como $\phi(B)\psi(B) = 1$ entonces los coeficientes ψ_j's pueden ser obtenidos en función de los ϕ_j's. Una condición para que X_t sea estacionario, es que todas las raíces de $\phi(B) = 0$ caen fuera del círculo unitario, o equivalentemente que

$|\phi_j| < 1$ para $j = 1, 2, \ldots, p$, esto es que los valores ϕ_j caigan dentro del círculo unitario [Box Jenkins y Reinsel, 1994].

Bajo el supuesto de estacionariedad y multiplicando (2.4) por X_{t-h} y tomando esperanzas se tiene:

$$\sigma_X^2 = \gamma(0) = \frac{\sigma^2}{1 - \phi_1\rho_1 - \phi_2\rho_2 - \ldots - \phi_p\rho_p} \qquad \text{Para } h=0 \qquad \textbf{(2.8)}$$

$$\gamma(h) = \phi_1\gamma(h-1) + \phi_2\gamma(h-2) + \ldots + \phi_p\gamma(h-p) \qquad \text{Para } h > 0 \qquad \textbf{(2.9)}$$

Que son la varianza y la función de autocovarianza del procesos *AR(p)*, respectivamente. De la misma manera para obtener la función de autocorrelación del proceso *AR(p)*, basta dividir todos los elementos de (2.9) por $\gamma(0)$, esto es:

$$\rho(h) = \phi_1\rho(h-1) + \phi_2\rho(h-2) + \ldots + \phi_p\rho(h-p) \qquad \textbf{(2.10)}$$

Tomando $\rho_0, \rho_1, \rho_2, \ldots, \rho_{p-1}$ como condiciones iniciales, determinadas a partir de los coeficientes $\phi_1, \phi_2, \ldots, \phi_p$, la solución de la ecuación (2.10) permite calcular los valores de $\rho(h)$, para $h \geq p$.

Particularizando (2.10) para $h = 1, 2, \ldots, p$ se obtiene el sistema de ecuaciones de *Yule-Walker*:

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2\rho_1 + \ldots + \phi_p\rho_{p-1} \\
\rho_2 &= \phi_1\rho_1 + \phi_2 + \ldots + \phi_p\rho_{p-2} \\
&\vdots \\
\rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \ldots + \phi_p
\end{aligned} \qquad \textbf{(2.11)}$$

Resolviendo el anterior sistema resulta que:

$$\phi_p = \mathrm{P}_p^{-1}\rho_p \qquad \textbf{(2.12)}$$

Donde $\mathrm{P}_p = [\rho_{ij}]$ con $\rho_{ij} = \rho_{|i-j|},\ i, j = 1, 2, \ldots, p$, $\phi_p = (\phi_1, \phi_2, \ldots, \phi_p)'$, $\rho_p = (\rho_1, \rho_2, \ldots, \rho_p)'$.

La ecuación (2.12) puede ser utilizada para obtener estimadores de los parámetros ϕ_j's, sustituyendo las funciones de autocorrelación por sus estimativas. Estos estimadores usualmente son llamados estimadores de *Yule-Walker* . Una solución general de (2.10), nos permite concluir que la función de autocorrelación de un proceso autorregresivo de orden *p* es una mezcla de funciones exponenciales y senoides amortiguadas.

2.6.3 Procesos de Medias Móviles MA(q)

Decimos que $\{X_t, t \in \mathrm{Z}\}$ es un proceso de medias móviles de orden *q*, se denota por $\{X_t\} \sim MA(q)$, si satisface la ecuación de diferencias:

$$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q} \qquad \textbf{(2.13)}$$

Donde $\mu, \theta_1, \theta_2, \ldots, \theta_q$ son parámetros del modelo y $\{\varepsilon_t\} \sim RB(0, \sigma^2)$. X_t es estacionario con media $E(X_t) = \mu$, y como ε_t son no correlacionados, podemos obtener fácilmente la varianza del proceso:

$$\sigma_X^2 = \sigma^2(1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2) \qquad \textbf{(2.14)}$$

Sin perder la generalidad supongamos $\mu = 0$, entonces la función de autocovarianza del proceso es:

$$\gamma(h)=\begin{cases}\sigma^2\sum_{j=0}^{q-|h|}\theta_j\theta_{j+|h|}, & \text{si} \quad |h|\leq q \\ 0, & \text{si} \quad |h|>q\end{cases} \tag{2.15}$$

Donde $\theta_0=1$, consecutivamente se puede obtener la función de autocorrelación $\rho(h)$ como la razón de $\gamma(h)/\sigma_X^2$, en particular $\rho(h)=0, \quad \text{si} \ |h|>q$. De manera general el proceso (2.13) puede ser escrito como:

$$X_t=\theta(B)\varepsilon_t \tag{2.16}$$

Donde $\theta(B)=1+\theta_1B+\theta_2B^2+\ldots+\theta_qB^q$ es el operador de media móvil o polinomio de media móvil de orden *q*, al proponer una solución para (2.16) del siguiente modo:

$X_t=\theta(B)\varepsilon_t \quad \Rightarrow \theta(B)^{-1}X_t=\varepsilon_t \quad \Rightarrow \pi(B)X_t=\varepsilon_t$

Así la solución es:

$$\sum_{j=0}^{\infty}\pi_jX_{t-j}=\varepsilon_t, \quad \pi_0=1 \tag{2.17}$$

Donde $\pi(B)=1+\pi_1B+\pi_2B^2+\ldots$, tal que $\sum_{j=0}^{\infty}|\pi_j|<\infty$, de modo que $\pi(B)=\theta(B)^{-1}$, por tanto los coeficientes π_j pueden ser obtenidos mediante $\pi(B)\theta(B)=1$, y para satisfacer la condición de invertibilidad, necesitamos que todas las raíces de $\theta(B)=0$ deben estar fuera del círculo unitario.

2.6.4 Procesos Autorregresivos de Medias Móviles ARMA(p,q)

Decimos que $\{X_t, t \in \mathrm{Z}\}$ es un proceso autorregresivo de medias móviles de orden (p,q) y se denota por $\{X_t\} \sim ARMA(p,q)$, si para cada $t \in \mathbb{Z}$:

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \ldots + \phi_p(X_{t-p} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q} \quad \textbf{(2.18)}$$

Donde $\mu, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$ son parámetros del modelo, y $\{\varepsilon_t\} \sim RB(0, \sigma^2)$, Usando los operadores autorregresivo y de media móvil, definidos anteriormente, podemos escribir (2.18) de la forma siguiente:

$$\phi(B)\tilde{X}_t = \theta(B)\varepsilon_t \quad \textbf{(2.19)}$$

Donde $\tilde{X}_t = X_t - \mu$, en adelante se asume que $\mu = 0$; para un proceso *ARMA(p,q)* genérico una condición de estacionariedad es la misma que para los procesos *AR(p)*, esto es, que las raíces de $\phi(B) = 0$ deben estar fuera del círculo unitario, del mismo modo una condición de invertibilidad es la misma que para los procesos *MA(q)*, esto es, que las raíces de $\theta(B) = 0$, deben estar fuera del círculo unitario.

Considerando $\mu = 0$, luego multiplicando (2.18), por X_{t-h}, y tomando esperanzas obtenemos la función de autocovarianza:

$$\gamma(h) = \phi_1\gamma(h-1) + \ldots + \phi_p\gamma(h-p) + \gamma_{X\varepsilon}(h) + \theta_1\gamma_{X\varepsilon}(h-1) + \ldots + \theta_q\gamma_{X\varepsilon}(h-q) \quad \textbf{(2.20)}$$

Donde $\gamma_{X\varepsilon}(h)$, es la covarianza cruzada de X_t y ε_t, definida por $\gamma_{X\varepsilon}(h) = E(\varepsilon_t X_{t-h})$, como X_{t-h} sólo depende de los choques ε_t ocurridos hasta el instante *t-h*, tenemos que esta covarianza cruzada sólo es diferente de cero para $h \leq 0$, luego:

$$\gamma(h) = \phi_1\gamma(h-1) + \phi_2\gamma(h-2) + \ldots + \phi_p\gamma(h-p), \qquad h > q \quad \textbf{(2.21)}$$

Esto es que las autocovarianzas y por tanto las autocorrelaciones, de retardos $1,2,\ldots,q$, son afectados por los parámetros de medias móviles, pero para $h>q$, tienen el mismo comportamiento que los modelos autorregresivos, una enumeración explícita de la función de autocorrelación para un proceso ARMA(p,q) se puede encontrar en Ramírez y Sastri (1997), en esta publicación se muestra que la función de autocorrelación para un proceso ARMA(q,q) es la función de los θ', ϕ' y σ^2.

2.7 Procesos Lineales no Estacionarios

La mayoría de series de tiempo presentan una no estacionariedad en media y algunas una no estacionariedad en varianza; esto es, varían sobre translaciones del tiempo. Una de las estrategias para expresar un modelo no estacionario, consiste en tomar diferencias regulares, y/o diferencias estacionales. En otras ocasiones se requieren de transformaciones no lineales sobre la serie observada, y algunas veces las dos operaciones juntas, siendo el objetivo principal obtener una serie estacionaria que permita la estimación de los parámetros basado en una realización del proceso. Dentro de esta clase de modelos se encuentran los procesos autoregresivos integrados de medias móviles, $ARIMA(p,d,q)$ y los procesos estacionales autoregresivos integrados de media móviles $SARIMA(p,d,q)(P,D,Q)_S$.

2.7.1 Proceso Autorregresivo Integrado y de Media Móvil ARIMA(p,d,q)

Para convertir una serie de tiempo no estacionaria en una serie estacionaria, típicamente se utiliza el operador ordinario de diferencias

Definición.- Si $\Delta^d X_t$ es un proceso estacionario, decimos que X_t es integrado de orden *d* y lo denotamos por $X_t \sim I(d)$, donde Δ^d, es el operador de diferencias regulares de orden *d*, es tal que : $\Delta^d X_t = (1-B)^d X_t$, $d \in \mathbb{Z}_0^+$,

Definición.- Si $\Delta^d X_t$ sigue un proceso $ARMA(p,q)$, decimos que X_t sigue un modelo $ARIMA(p,d,q)$, esto es:

$$\phi(B)\Delta^d(X_t - \mu) = \theta(B)\varepsilon_t \qquad \textbf{(2.22)}$$

Sin perder la generalidad, consideramos $\mu = 0$, así (2.22), se reduce a

$$\phi(B)\Delta^d X_t = \theta(B)\varepsilon_t \qquad \textbf{(2.23)}$$

De modo equivalente (2.23), se puede escribir como $\phi(B)W_t = \theta(B)\varepsilon_t$, con $W_t = \Delta^d X_t$; observamos que $W_t = \Delta^d X_t \Leftrightarrow X_t = S^d W_t$, donde *S* es el operador Suma o Integral, tal que $S = (1-B)^{-1} = \Delta^{-1}$. Así X_t puede ser obtenido sumando o integrando el proceso estacionario W_t, *d* veces.

2.7.2 Proceso Estacional Autorregresivo Integrado y de Media Móvil SARIMA(p,d,q)(P,D,Q)$_S$

Cuando una serie de tiempo en estudio tiene intervalos de observación menores a un año, entonces es frecuente que estas tengan variaciones ó patrones sistemáticos cada cierto periodo, estas variaciones sistemáticas inferiores a un año por ejemplo semestral, mensual, diario, etc. Deben ser captadas en los llamados Factores Estacionales, dentro de la estructura del modelo a construirse.

Las series de tiempo estacionales pueden ser de dos tipos:

- Aditivas.
- Multiplicativas.

Y al mismo tiempo cada una de estas series pueden ser estacionarias o no estacionarias.

Usualmente se presentan con mayor frecuencia los modelos multiplicativos comparados con los modelos aditivos, de esta manera se combinan términos ordinarios del proceso *ARMA* y términos estacionales, así como diferencias regulares y diferencias estacionales; por lo tanto la estructura general de un modelo $SARIMA(p,d,q)(P,D,Q)_S$, es:

$$\phi(B)\Phi(B^s)\Delta^d\Delta_s^D(X_t-\mu)=\theta(B)\Theta(B^s)\varepsilon_t \qquad \textbf{(2.24)}$$

Donde:

$\phi(B)=1-\phi_1B-\phi_2B^2-\ldots-\phi_pB^p$, es el polinomio autorregresivo estacionario de orden p.

$\theta(B)=1+\theta_1B+\theta_2B^2+\ldots+\theta_qB^q$, es el polinomio de medias móviles invertible de orden q.

$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{sP}$, es el polinomio autorregresivo estacional de orden P, estacionario.

$\Theta(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{sQ}$, es el polinomio de medias móviles estacional de orden Q, invertible.

$\Delta^d = (1 - B)^d$, es el operador de diferencias regulares, con $d \in \mathbb{Z}_0^+$ indicando el número de diferencias regulares.

$\Delta_s^D = (1 - B^s)^D$, es el operador de diferencias estacionales, con $D \in \mathbb{Z}_0^+$ indicando el número de diferencias estacionales.

s, es el periodo del proceso; y μ es la media del proceso.

Sin perder la generalidad consideramos $\mu = 0$ y $Y_t = (1 - B)^d (1 - B^s)^D X_t$, entonces (2.24) queda expresado como:

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)\varepsilon_t \qquad \textbf{(2.25)}$$

Así Y_t constituye un proceso $ARMA(p + sP, q + sQ)$.

2.8 Modelamiento de un Proceso ARIMA(p,d,q)

En esta parte trataremos de modelar series temporales por medio de procesos autorregresivos integrados de medias móviles $ARIMA(p, d, q)$.

Dado que los procesos estocásticos $ARIMA(p, d, q)$ pertenecen a una familia amplia de procesos lineales, se hace necesario el uso de una estrategia de construcción de modelos, el cual esta constituido de las siguientes etapas:

- Identificación de la Estructura,
- Estimación de los Parámetros,
- Diagnóstico del Modelo.

Este ciclo iterativo fue descrito y popularizado por *Box and Jenkins* (1976), para el estudio de los modelos de series de tiempo que siguen un comportamiento de los procesos $ARIMA(p,d,q)$, y generalizados para los modelos $SARIMA(p,d,q)(P,D,Q)_S$.

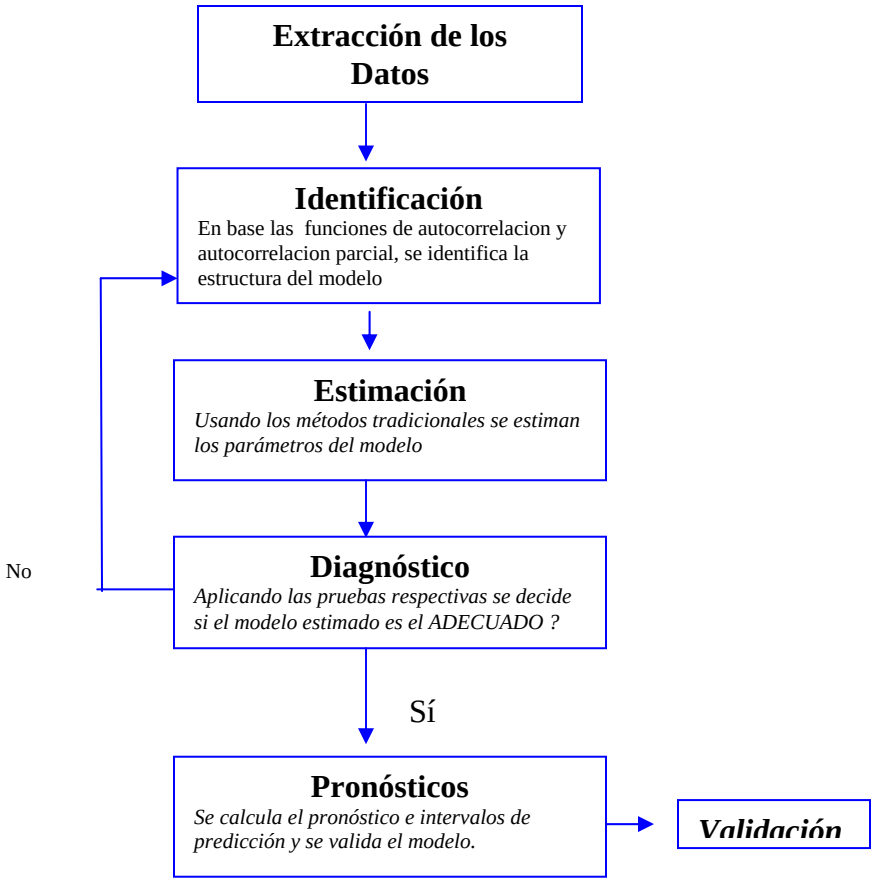


Figura 2.3: Ciclo iterativo básico de *Box y Jenkins*

Cada etapa tiene la respectiva importancia, pero la etapa crucial es la de identificación, ya que requiere mucha experiencia en el análisis exploratorio de los datos, y por lo general se buscara identificar modelos parsimoniosos, es decir modelos con el menor número de

parámetros. Una vez que el modelaje de las tres etapas ha concluido, se procede a realizar pronósticos con el mejor modelo que se ha identificado.

Dentro de la clase amplia de procesos lineales, los procesos que están en equilibrio; esto es, mantienen la misma distribución de probabilidades en todo el intervalo de observación, son los llamados procesos estacionarios, y bajo la literatura de *Box y Jenkins* se denominan, Procesos $ARMA(p,q)$

2.8.1 Identificación de un Modelo ARIMA(p,d,q)

La identificación de un modelo $ARIMA(p,d,q)$ en particular, para el modelamiento de nuestra serie de tiempo, es la fase más crítica del proceso iterativo descrito, así habiendo formulado una clase amplia de procesos estocásticos para las series de tiempo no estacionarias, ahora se deseara fijar o seleccionar valores de d y D diferencias regular y estacional respectivamente, que hacen que la serie de tiempo se vuelva estacionaria, y consecuentemente proponer valores de p,q,P,Q . Sin embargo se recomienda seguir primero el análisis de estabilización en la varianza, mediante la transformación de *Box y Cox* La cual se puede expresar como:

$$f_\lambda(X_t)=\begin{cases}\dfrac{X_t^{\lambda}-1}{\lambda}, & \text{si} \quad X_t \geq 0, \lambda > 0 \\ \ln(X_t), & \text{si} \quad X_t > 0, \lambda = 0\end{cases} \qquad \textbf{(2.26)}$$

Donde λ es el parámetro de transformación, de esta forma se verifica si existe la necesidad de transformar la serie original con el objetivo de estabilizar su varianza. Consecutivamente se toman diferencias regulares o estacionales, según sea el caso, tantas veces sea necesario

para obtener una serie estacionaria, de modo que el proceso $\Delta^d \Delta_s^D X_t$ sea reducido a un $ARMA(p+sP, q+sQ)$ el numero de diferencias d y D, necesarias para que el proceso se torne estacionario es alcanzado cuando la función de autocorrelación muestral y la función de autocorrelación parcial muestral, de $Y_t = \Delta^d \Delta_s^D X_t$, decrece rápidamente a cero. En las siguientes fases se considera únicamente en caso regular; ya que el caso estacional es una simple generalización del estudio $ARIMA(p,d,q)$.

La identificación en particular de un modelo *ARIMA* es hecha principalmente en base a las autocorrelaciones y autocorrelaciones parciales muestrales, que se espera representen adecuadamente las respectivas cantidades teóricas, que son desconocidas. Sabemos que la función de autocorrelación $\rho(h)$ es estimada por :

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = r_h = \frac{c_h}{c_0} \quad \text{con } h = 0,1,\ldots T-1 \qquad \textbf{(2.27)}$$

Donde $\hat{\gamma}(h) = c_h$ es la estimación de la función de autocovarianza $\gamma(h)$.

$$c_h = \frac{1}{T}\sum_{j=1}^{T-h}(x_{j+h} - \overline{x})(x_j - \overline{x}), \quad h = 0,1,\ldots T-1 \qquad \textbf{(2.28)}$$

Donde $\overline{x} = (1/T)\sum_{j=1}^{T} x_j$, es la media muestral de la serie, una expresión aproximada para la varianza de r_h, para un proceso estacionario normal esta dado por:

$$\text{var}(r_h) \simeq \frac{1}{T}\sum_{j=-\infty}^{\infty}\left[\rho_j^2 + \rho_{j+h}\rho_{j-h} - 4\rho_h\rho_j\rho_{j-h} + 2\rho_j^2\rho_h^2\right] \qquad \textbf{(2.29)}$$

Para un proceso $MA(q)$ en que las autocorrelaciones son nulas para $j > q$, todos los términos del lado derecho de (2.29) se anulan para $h > q$, excepto el primero, obteniéndose:

$$\text{var}(r_h) \simeq \frac{1}{T}\left[1 + 2\sum_{j=1}^{q} \rho_j^2\right], \quad h > q \qquad \textbf{(2.30)}$$

Como desconocemos las autocorrelaciones ρ_j las sustituimos por r_j, obteniendo el estimado de (2.30) como:

$$\widehat{\sigma^2}(r_h) \simeq \frac{1}{T}\left[1 + 2\sum_{j=1}^{q} r_j^2\right] \quad h > q \qquad \textbf{(2.31)}$$

Para *T* suficientemente grande y sobre la hipótesis que $\rho(h) = 0$, para $h > q$, la distribución de r_j es aproximadamente normal, con media igual a cero y varianza dada por (2.30), *Jenkins y Watts* (1968). Así es posible construir un intervalo de confianza aproximado para las autocorrelaciones dado por:

$$r_h \pm t_\gamma \hat{\sigma}(r_h) \qquad \textbf{(2.32)}$$

Donde t_γ es el valor de la estadística *t* de Student con *T*-1 grados de libertad. En la practica usualmente se asume que $t_\gamma = 2$ correspondiendo a $\gamma = 0.95$ aproximadamente, de modo que podemos considerad $\rho(h)$ como significativamente diferente de cero si $|r_h| > 2\hat{\sigma}(r_h)$, para $h > q$. De esta forma podemos caracterizar los diferentes procesos autorregresivos, medias móviles y la combinación de estos dos, y posteriormente identificar sus respectivos órdenes del siguiente modo:

I. Un proceso *AR(p)* tiene la función de autocorrelación que decae de acuerdo a una exponencial ó senosoides amortiguadas, extensas.

II. Un proceso *MA(q)* tiene la función de autocorrelación finita y presenta un corte en el orden del modelo *q*, es decir toma valores de cero para $h > q$.

III. Un proceso *ARMA(p,q)* tiene la función de autocorrelación que decae de acuerdo a una exponencial ó senosoides amortiguadas extensas en el retardo *q-p.*

Análogamente *Box Jenkins y Reinsel* (1994), proponen la utilización de otro instrumento para facilitar el procedimiento de identificación, la función de autocorrelación parcial, denotemos por ϕ_{pj} el *j-ésimo* coeficiente de un modelo *AR(p)* de tal modo que ϕ_{pp} sea el último coeficiente, por (2.10), y por la ecuaciones de *Yule Walker* (2.11), resolviendo tenemos:

$$\phi_{11} = \rho_1$$

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}}$$

$$\phi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho 1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}}$$

En general:

$$\phi_{pp} = \frac{|\mathrm{P}_p^*|}{|\mathrm{P}_p|} \qquad \textbf{(2.33)}$$

Donde P_p es la matriz de autocorrelaciones, y P_p^* es la matriz P_p con la última columna sustituida por el vector de autocorrelaciones. La cantidad ϕ_{pp} que ésta en función de *p*, es llamada función de autocorrelacion parcial, esto es: $Corr(X_{t+h}, X_t \mid X_{t+1}, \ldots, X_{t+h-1})$, que es la correlación entre X_{t+h} y X_t después de que sus dependencias lineales sobre las variables

que intervienen $X_{t+1}, X_{t+2}, \ldots, X_{t+h-1}$ haya sido removida. Una forma de encontrar las estimaciones de la función de autocorrelación parcial, es sustituyendo en (2.33), los ϕ_{pj}, por sus estimativas $\hat{\phi}_{pj}$, ρ_j por sus estimativas r_j, obteniendo así : $\hat{\phi}_{pp} = \frac{|\hat{P}_p^*|}{|\hat{P}_p|}$.

Si el número de observaciones *n*, es suficientemente grande, $\hat{\phi}_{pp}$ tiene una distribución aproximadamente normal, para un proceso *AR(p)*, con varianza dada por:

$$\text{var}(\hat{\phi}_{pp}) \approx \frac{1}{T} \quad , p \geq p+1 \qquad \textbf{(2.34)}$$

De esta forma podemos caracterizar los diferentes procesos autorregresivos, medias móviles y la combinación de estos dos, y posteriormente identificar sus respectivos ordenes del siguiente modo:

I. Un proceso *AR(p)* tiene la función de autocorrelación parcial $\phi_{kk} \neq 0$, para $k \leq p$ y $\phi_{kk} = 0$ para $k > p$.

II. Un proceso *MA(q)* tiene la función de autocorrelación parcial similar a la función de autocorrelación de un *AR(p)*, dominada especialmente por comportamientos exponenciales y senoidales amortiguadas.

III. Un proceso *ARMA(p,q)* tiene el comportamiento de la función de autocorrelación parcial similar a la función de autocorrelación parcial de un media móvil puro [Morettin, 2004].

Así mismo para T grande y sobre la hipótesis que un proceso sea *AR(p)*, $\hat{\phi}_{jj}$ tiene distribución aproximadamente normal, con media cero y varianza dado por (2.34), de modo que consideramos ϕ_{jj} significativamente diferente de cero, si:

$$|\hat{\phi}_{jj}| > \frac{1.96}{\sqrt{T}} \approx \frac{2}{\sqrt{T}}, \text{ para } j > p \qquad \textbf{(2.35)}$$

Así queda determinado los intervalos de confianza para los estimadores de la función de autocorrelacion parcial.

Existen formas alternativas de identificación de modelos *ARMA(p,q)*, la idea es escoger los órdenes *p* y *q* que minimicen una cantidad:

$$P(p,q) = \ln \hat{\sigma}^2_{p,q} + (p+q)\frac{C(n)}{n} \qquad \textbf{(2.36)}$$

Donde $\hat{\sigma}^2_{p,q}$ es el estimado de la varianza residual obtenida ajustando un modelo *ARMA(p,q)*, $C(n)$ es una función de tamaño de la serie, la cantidad $(p+q)\frac{C(n)}{n}$, es denominado término penalizador, que aumenta cuando el número de parámetros aumenta y cuando la varianza residual disminuye. Existen algunos procedimientos de identificación que minimizan las funciones penalizadotas particulares así tenemos:

- **Criterio de Información Akaike (AIC)** *Akaike*, sugiere escoger un modelo cuyos órdenes *p* y *q* minimizan: $\text{AIC}(p,q) = \ln \hat{\sigma}^2_{p,q} + \frac{2(p+q)}{n}$, donde $\hat{\sigma}^2_{p,q}$ es el estimador de máxima verosimilitud de la varianza residual bajo el modelo *ARMA(p,q)*. Existen correcciones para mejorar el comportamiento del AIC, en el sentido de disminuir la probabilidad de seleccionar un orden mayor que el verdadero, [Morettin, 2004].

- **Criterio de Información de Akaike Corregido (AICC)** *Hurvich and Tsai* (1989) proponen una corrección para el AIC que en el caso de un proceso autorregresivo esta dado por: $\mathrm{AICC}(p) = \mathrm{AIC}(p) + \frac{2(p+1)(p+2)}{n-p+2}$ con $0 \le p \le p_{\max}$. Este resultado se puede generalizar para un modelo autorregresivo estacional [Morettin, 2004].
- **Criterio de Información Bayesiano (BIC)** Sugiere minimizar el criterio de información bayesiano dado por: $\mathrm{BIC}(p,q) = \ln \hat{\sigma}^2_{p,q} + (p+q)\frac{\ln n}{n}$ [Morettin, 2004].

2.8.2 Estimación de un Modelo ARIMA(p,d,q)

Teniendo identificado un modelo tentativo para nuestra serie temporal, el siguiente paso es estimar sus parámetros, los métodos de mínimos cuadrados y máxima verosimilitud pueden ser usados para este propósito.

Para un modelo general estacionario *ARMA(p,q),* dado por:

$X_t - \mu = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \ldots + \phi_p(X_{t-p} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q}$ donde

$\mu, \boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)', \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)'$ son parámetros del modelo, y $\{\varepsilon_t\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ ruido blanco, la función de densidad de probabilidad conjunta de $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n)'$, esta dado por:

$$f(\boldsymbol{\varepsilon} \mid \boldsymbol{\phi}, \mu, \boldsymbol{\theta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{t=1}^{n} \varepsilon_t^2 \right] \qquad \textbf{(2.37)}$$

Reescribiendo el modelo *ARMA(p,q)* como

$\varepsilon_t = (X_t - \mu) - \phi_1(X_{t-1} - \mu) - \phi_2(X_{t-2} - \mu) - \ldots - \phi_p(X_{t-p} - \mu) + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q}$,

podemos escribir la función de verosimilitud de los parámetros $(\boldsymbol{\phi}, \mu, \boldsymbol{\theta}, \sigma^2)$. Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$, y asumimos que las condiciones iniciales $\boldsymbol{X}_* = (X_{1-p}, \ldots, X_{-1}, X_0)'$, y $\boldsymbol{\varepsilon}_* = (\varepsilon_{1-q}, \ldots, \varepsilon_{-1}, \varepsilon_0)'$, la función condicional Log-verosimilitud ésta dada por:

$$\ln L_*(\boldsymbol{\phi}, \mu, \boldsymbol{\theta}, \sigma^2) \approx \frac{n}{2} \ln 2\pi\sigma^2 - \frac{S_*(\boldsymbol{\phi}, \mu, \boldsymbol{\theta})}{2\sigma^2} \qquad \textbf{(2.38)}$$

Donde $S_*(\boldsymbol{\phi}, \mu, \boldsymbol{\theta}) = \sum_{t=1}^{n} \varepsilon_t^2(\boldsymbol{\phi}, \mu, \boldsymbol{\theta} \mid \boldsymbol{X}_*, \boldsymbol{\varepsilon}_*, \boldsymbol{X})$, es la función condicional de la suma de cuadrados. Las cantidades $\hat{\mu}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}$ las cuales maximizan la ecuación (2.38) son llamados estimadores de máxima verosimilitud condicionales. Desde que $\ln L_*(\boldsymbol{\phi}, \mu, \boldsymbol{\theta}, \sigma^2)$ envuelve la data sólamente a través de $S_*(\boldsymbol{\phi}, \mu, \boldsymbol{\theta})$, estos estimadores son los mismos que los estimadores de mínimos cuadrados condicionales, obtenidos de minimizar la función suma de cuadrados condicionales $S_*(\boldsymbol{\phi}, \mu, \boldsymbol{\theta})$.

Hay pocas alternativas para especificar los valores iniciales $\boldsymbol{X}_*$ y $\boldsymbol{\varepsilon}_*$. Basados en la presunción que $\{X_t\}$ es estacionario y $\{\varepsilon_t\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, podemos reemplazar el X_t desconocido por la media muestral $\bar{X}$, y el ε_t desconocido por su valor esperado cero. Para el modelo *ARMA(p,q)* también podemos asumir que $\varepsilon_p = \varepsilon_{p-1} = \ldots = \varepsilon_{p+1-q} = 0$ y calcular ε_t para $t \geq (p+1)$, así la función condicional de suma de cuadrados se simplifica en: $S_*(\boldsymbol{\phi}, \mu, \boldsymbol{\theta}) = \sum_{t=p+1}^{n} \varepsilon_t^2(\boldsymbol{\phi}, \mu, \boldsymbol{\theta} \mid \boldsymbol{X})$, el cual es usado por la mayoría de programas de

computadoras. Luego de obtener los estimadores de los parámetros $\hat{\mu}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}$, el estimador de σ^2, esta dado por: $\hat{\sigma}^2 = \frac{S_*(\hat{\boldsymbol{\phi}}, \hat{\mu}, \hat{\boldsymbol{\theta}})}{n-(2p+q+1)}$. Para tratar los modelos $SARIMA(p,d,q)(P,D,Q)_S$, la estimación de los parámetros se comporta de manera analógica a lo presentado, únicamente ingresando en el modelo parámetros estacionales [Morettin, 2004].

2.8.3 Diagnóstico de un Modelo ARIMA(p,d,q)

Luego de la estimación de los parámetros la tarea se centra en verificar si el modelo representa ó no adecuadamente los datos. Una técnica que puede ser utilizada es el **superajustamiento** que consiste en estimar un modelo con parámetros extras y examinamos si estos son significativos, y también si, su inclusión disminuye significativamente la varianza residual. Este método está basado en la experiencia. Así mismo existen técnicas alternativas como las que se describe a continuación:

- **Test de Autocorrelación Residual.** luego de estimar los parámetros, los $\hat{\varepsilon}_t = \hat{\theta}^{-1}(B)\hat{\Theta}^{-1}(B)\phi(B)\Phi(B)Y_t$, son llamados residuos, si el modelo es adecuado, los $\hat{\varepsilon}_t$ deben estar próximos a ε_t, y por tanto deben ser aproximadamente no correlacionados; si indicamos que $\hat{r}_k$, son las autocorrelaciones de los residuos $\hat{\varepsilon}_t$, deberíamos tener $\hat{r}_k \simeq 0$, en particular $\hat{r}_k \sim N(0, \frac{1}{n})$, donde *n* es el número efectivo de observaciones. De este modo, una comparación de $\hat{r}_k$ con los limites aproximados

$\pm\frac{2}{\sqrt{n}}$, muestran una indicación general de un posible comportamiento de un ruido blanco en ε_t.

- **Test de Ljung-Box (1978).** Es una prueba para las autocorrelaciones de los residuos estimados, que a pesar de no detectar ciertas especificaciones en el comportamiento de ruido blanco, puede indicar que esos valores son muy altos; esta prueba es una variante de la versión original del test de *Box and Pierce*. Si el modelo es apropiado, la estadística de prueba es: $Q(K) = n(n+2)\sum_{k=1}^{K}\frac{\hat{r}_k^2}{(n-k)}$, tiene una distribución Chi-cuadrado con $K-p-q$ grados de libertad $\left(\chi^2_{K-p-q}\right)$, la hipótesis de ruido blanco para los residuos es rechazada para valores altos de $Q(\cdot)$. La mayoría de programas de computadoras utilizan este test.

Finalmente si se cumple con los pasos anteriores realizando un análisis exhaustivo en cada uno de ellos, el próximo paso es poder realizar pronósticos [Uriel ,1985].

2.8.4 Predicciones con Modelos ARIMA(p,d,q)

Queremos pronosticar valores X_{T+h}, teniendo observaciones hasta el instante *T*, usando un modelo *ARIMA*; sea $\varphi(B) = \phi(B)\Delta^d = 1-\varphi_1 B-\varphi_2 B^2-\ldots-\varphi_{p+d}B^{p+d}$, llamemos $\hat{X}_T(h)$ a la predicción de X_{T+h}, con origen en *T* y *h* pasos adelante. Se puede probar que la predicción con error cuadrático medio mínimo esta dada por la esperanza condicional de X_{T+h} dado el pasado $X_T, X_{T-1}, X_{T-2},\ldots$, esto es:

$$\hat{X}_T(h) = E(\varphi_1 X_{T+h-1} + \ldots + \varphi_{p+d} X_{T+h-p-d} + \theta_0 + a_{T+h} + \theta_1 a_{T+h-1} + \ldots + \theta_q a_{T+h-q} \mid X_T, X_{T-1}, \ldots) \quad \textbf{(2.39)}$$

Donde $a_{T+h}, a_{T+h-1} \ldots, a_{T+h-q}$, representan los residuales de la serie diferenciada [Morettin, 2004]. Para calcular las predicciones usamos los hechos:

a) $E\left(X_{T+j} \mid X_T, X_{T-1}, \ldots\right) = \begin{cases} X_{T+j}, & \text{si } j \le 0 \\ \hat{X}_T(j), & \text{si } j > 0 \end{cases}$

b) $E\left(\varepsilon_{T+j} \mid X_T, X_{T-1}, \ldots\right) = \begin{cases} \varepsilon_{T+j}, & \text{si } j \le 0 \\ 0, & \text{si } j > 0 \end{cases}$

Luego para calcular las predicciones tenemos que

- Sustituir las esperanzas pasadas $(j \le 0)$ por los valores conocidos X_{T+j} y ε_{T+j}
- Sustituir las esperanzas futuras $(j > 0)$ por las predicciones $\hat{X}_T(j)$ y 0.

Escribiendo el modelo en la forma infinita de medias móviles, se prueba que el error predicción está dado por:

$$e_T(h) = X_{T+h} - \hat{X}_T(h) = \varepsilon_{T+h} + \psi_1 \varepsilon_{T+h-1} + \ldots + \psi_{h-1} \varepsilon_{T+1} \quad \textbf{(2.40)}$$

Donde los coeficientes ψ_j provienen de $\psi(B) = \varphi^{-1}(B)\theta(B)$, así la varianza del error de predicción está dada por:

$$V(h) = \operatorname{var}\left(e_T(h)\right) = \sigma^2(1 + \psi_1^2 + \ldots + \psi_{h-1}^2) \quad \textbf{(2.41)}$$

Como $e_T(1) = X_{T+1} - \hat{X}_T(1) = \varepsilon_{T+1}$, los errores de predicción a un paso adelante son no correlacionados, suponiendo que todos los parámetros del modelo son conocidos. En la

práctica se utiliza el modelo estimado para realizar las predicciones, así (2.41) es $\hat{V}(h) = \hat{\sigma}^2(1 + \sum_{j=1}^{h-1} \hat{\psi}_j^2)$.

Finalmente para determinar un intervalo de confianza para X_{t+h} será necesario hacer una suposición adicional para los residuos, esto es: $E(\varepsilon_t) = 0$, $\text{var}(\varepsilon_t) = \sigma^2$, para todo t y $E(\varepsilon_t a_s) = 0$, $t \neq s$, además $\varepsilon_t \sim N(0, \sigma^2)$, para cada t. Se sigue que dados los valores pasados y presentes de la serie $X_T, X_{T-1}, X_{T-2}, \ldots$, la distribución condicional de X_{T+h} será $N(\hat{X}_T(h), V(h))$. Así $Z = \dfrac{X_{T+h} - \hat{X}_T(h)}{[V(h)]^{1/2}} \sim N(0,1)$; fijando un coeficiente de confianza γ, podemos encontrar un valor $Z_{\alpha/2}$, tal que $P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = \gamma$, así $\hat{X}_T(h) - Z_{\alpha/2}[V(h)]^{1/2} \leq X_{T+h} \leq \hat{X}_T(h) + Z_{\alpha/2}[V(h)]^{1/2}$, es el intervalo de confianza para la predicción X_{T+h}. Sustituyendo $V(h)$ por su estimación $\hat{V}(h)$, tenemos:

$$\hat{X}_T(h) - Z_{\alpha/2}\hat{\sigma}\left[1 + \sum_{j=1}^{h-1} \hat{\psi}_j^2\right]^{1/2} \leq X_{T+h} \leq \hat{X}_T(h) + Z_{\alpha/2}\hat{\sigma}\left[1 + \sum_{j=1}^{h-1} \hat{\psi}_j^2\right]^{1/2} \qquad \textbf{(2.42)}$$

Así tenemos construido el intervalo de confianza para la predicción X_{T+h}.

3 METODOLOGIA BOOTSTRAP

3.1 Introducción

El *bootstrap* es un método computacional intensivo que provee respuestas a una gran clase de problemas de inferencia estadística. No asume estructuras estrictas sobre los procesos aleatorios que generan la data; su principio básico es simple en sus diferentes formas y aspectos, que intentan recrear la relación entre población y muestra, considerando la muestra como una personificación de la población y mediante un remuestreo de éste, generando la "muestra *bootstrap*" el cual sirve como un análogo de la muestra dada; obviamente el mecanismo de remuestreo debe ser adecuado de tal manera que la muestra con la remuestra reflejen la relación original de la población y la muestra. El uso de la terminología *bootstrap* deriva de la frase *to pull oneself up by one's bootstrap*. El *bootstrap* desde su creación por *Efron B.* (1979), se ha aplicado para resolver numerosos problemas estadísticos, llegando incluso a tener mejor rendimiento que la metodología existente, como en el caso de estimar los errores estándar de algunos estadísticos (media, mediana, etc.). En muchos problemas complejos donde las aproximaciones convencionales fallan, el bootstrap nos provee una respuesta satisfactoria. Sin embargo, no resuelve todos los problemas de inferencia estadística. En el presente trabajo se considera la aplicación de bootstrap a una clase de procesos dependientes que hace notar situaciones donde puede ser aplicado efectivamente diferentes tipos de métodos *bootstrap*.

Los métodos de remuestreo y el *bootstrap* típicamente aplican a problemas de inferencia estadística que envuelven parámetros de nivel-2 y nivel-alto [Lahiri, 2003], de un proceso determinado, esto es:

Sea la secuencia de variables aleatorias $X_1, X_2, \ldots$ con distribución conjunta de probabilidad P, supongamos que la secuencia puede ser modelada como una realización de las *n* primeras variables aleatorias $\{X_1, \ldots, X_n\} \equiv \chi_n$, además que $\theta = \theta(P)$ es el parámetro de interés, el cual depende de la distribución conjunta desconocida P, el problema es encontrar un estimador de θ basado en las observaciones χ_n. Uno de los métodos comunes para encontrar estimadores de θ son los basados en la teoría de máxima verosimilitud, técnica de momentos y métodos no parametricos, etc.

Supongamos que un estimador de θ es $\hat{\theta}_n$ basado en χ_n, ahora se ve la necesidad de saber cuan exacto es el estimador $\hat{\theta}_n$ (sesgo y eficiencia); denotemos con G_n a la distribución muestral del estimador centrado $\hat{\theta}_n - \theta$, por que la distribución conjunta de χ_n es desconocida, G_n típicamente también es desconocida, así las cantidades como el error cuadrático medio de $\hat{\theta}_n$, $\mathrm{MSE}(\hat{\theta}_n)$ y los cuantiles de $\hat{\theta}_n$, son cantidades poblacionalmente desconocidas basadas en la distribución muestral G_n. En el presente trabajo llamamos parámetros de nivel-1, a los parámetros como θ y parámetros de nivel-2 a los cuales relacionan la distribución muestral del estimador de un parámetro nivel-1, como es el caso del $\mathrm{MSE}(\hat{\theta}_n)$. Es así que el *bootstrap* y otros métodos de remuestreo pueden ser

considerados métodos generales para encontrar estimadores de los parámetros de nivel-2. En forma general para estimar parámetros de cualquier nivel, es necesario usar un número apropiado de iteraciones del *bootstrap*, o se puede aplicar combinaciones sucesivas de más de un método de remuestreo para parámetros de nivel-alto.

La técnica *bootstrap* es usualmente más sencilla para data de tipo independiente y tiene un desarrollo un tanto más complicado para datos dependientes, pero literalmente el funcionamiento es el mismo para los dos casos. A continuación se presenta una breve descripción del *bootstrap* para datos independientes. Supongamos que $X_1,\ldots,X_n$ son variables aleatorias independientes e idénticamente distribuidos (iid), con distribución de probabilidad común F, luego su distribución conjunta de probabilidad es $P_n = F^n$; el parámetro de nivel-1 θ, es una función de la distribución F, esto es, $\theta = \theta(F)$, sea $\hat{\theta}_n = t(X_1,\ldots,X_n)$ su estimador. Suponer que se está interesado en el error cuadrático medio (MSE) de $\hat{\theta}_n$, es claro que tanto el MSE($\hat{\theta}_n$) como la distribución del estimador centrado $\hat{\theta}_n - \theta$, dependen de F, el cual es por naturaleza desconocido. El *bootstrap* facilita la solución de este problema sin requerir del conocimiento pleno de la población, el primer paso envuelve la construcción de un estimador $\tilde{F}_n$ de F, de la muestra $X_1,\ldots,X_n$ la cual se presume ser representativa de la población. El siguiente paso envuelve la generación de variables aleatorias iid $X_1^*,\ldots,X_n^*$ de $\tilde{F}_n$ obviamente condicionado sobre la muestra $X_1,\ldots,X_n$, la cual ahora toma el rol de población para la versión del *bootstrap* original; así la versión *bootstrap* del estimador $\hat{\theta}_n$, es $\hat{\theta}_n^*$ que se obtiene de reemplazar $X_1,\ldots,X_n$ por

$X_1^*,\ldots,X_n^*$ y la versión *bootstrap* del parámetro de nivel-1 $\theta = \theta(F)$ esta dado por $\theta(\tilde{F}_n)$, Debe notarse que en estos cálculos se esta usando únicamente el conocimiento de la muestra $X_1,\ldots,X_n$. Para una elección razonable de $\tilde{F}_n$ la versión *bootstrap* reproduce exactamente esas características de la población y la muestra que determina la distribución muestral de las variables como $\hat{\theta}_n - \theta$. El *bootstrap* se puede utilizar para estimar un parámetro de nivel-2 relacionado con la distribución desconocida de $\hat{\theta}_n - \theta$. Específicamente el estimador *bootstrap* de la distribución muestral desconocida G_n de $\hat{\theta}_n - \theta$, esta dado por la distribución condicional $\hat{G}_n$ de la versión *bootstrap* $\theta_n^* - \theta(\hat{F}_n)$. Un estimador *bootstrap* de un parámetro de nivel-2, $\varphi_n \equiv \varphi(G_n)$, es de la forma funcional dado por $\hat{\varphi}_n \equiv \varphi(\hat{G}_n)$ como por ejemplo el estimador *bootstrap* del sesgo y MSE de $\hat{\theta}_n$, esto es los dos primeros momentos de $\hat{\theta}_n - \theta$:

$$\widehat{BIAS} = \int x\hat{G}_n(dx) = E_*(\theta_n^*) - \theta(\tilde{F}_n) \qquad \textbf{(3.1)}$$

$$\widehat{MSE} = \int x^2\hat{G}_n(dx) = E_*[(\theta_n^*) - \theta(\tilde{F}_n)]^2 \qquad \textbf{(3.2)}$$

Donde E_* denota la esperanza condicional dado $X_1,\ldots,X_n$; esto es para algún estimador $\hat{\theta}_n$ y no presupone ninguna forma especifica. La elección para $\tilde{F}_n$ usualmente es la función de distribución empírica; las variables $X_1^*,\ldots,X_n^*$ son extraídas con remplazamiento de la muestra $X_1,\ldots,X_n$.

3.2 Muestra aleatoria

Una muestra aleatoria es una colección de *n* unidades $u_1,\ldots,u_n$ seleccionadas de una población comprendida de *N* unidades $U_1,\ldots,U_N$, cada una con igual probabilidad de ser seleccionada; en realidad se seleccionan *n* números enteros $j_1,\ldots,j_n$ del 1 al *N* con probabilidad 1/*N*, estos números serán los índices de las unidades seleccionadas a la muestra. En un muestreo con remplazamiento las unidades extraídas se pueden repetir, caso contrario el muestreo será sin reemplazo. Uno usualmente tiene interés en las mediciones x_i de las unidades u_i, así $x=(x_1,\ldots,x_n)$ son las mediciones observadas en la muestra, provenientes de las mediciones de la población $X=(X_1,\ldots,X_n)$, por tanto x es muestra aleatoria de *X*.

3.3 Función de distribución de una variable aleatoria

Una variable aleatoria es una función que asume valores de acuerdo a los resultados de un experimento aleatorio, una variable aleatoria puede asumir valores discretos o continuos, de acuerdo al espacio muestral del experimento aleatorio.

La función de distribución F de una variable aleatoria *X*, está dada por

$$F_X(t)=P(X\le t), \quad \text{para } t\in\mathbb{R}. \tag{3.3}$$

Sea una variable aleatoria X con función de distribución F_X, entonces $x_1,\ldots,x_n$ se llama muestra aleatoria de F_X, si $F_{x_i}=F_X$ y las x_i son independientes entre sí, $\forall i$.

3.4 Función de Distribución Empírica

Sea la muestra aleatoria $x_1,\ldots,x_n$, observada de F, esto es $F \to (x_1,\ldots,x_n)$, la función de distribución empírica para la muestra aleatoria está definida por

$$\hat{F}_n(t) = \frac{\#\{x_i \le t\}}{n} \qquad \textbf{(3.4)}$$

Donde para cada t, $\hat{F}_n(t)$ es un estadístico, cuyo resultado es la frecuencia relativa de los valores muestrales que son menores o iguales que t. Teóricamente la función de distribución empírica goza de muchas propiedades una de las principales es que $E[\hat{F}_n(t)] = F(t)$, siguiendo este hecho se puede considerar a la función de distribución empírica como un estimador no paramétrico de la función de distribución de probabilidad de una variable aleatoria.

3.5 El Principio Plug-in

Es un método para obtener un estimador de un parámetro $\theta = t(F)$ donde F es una función de distribución conjunta de $X_1, X_2,\ldots,$ así el estimador *plug-in* o análogo es $\hat{\theta} = t(\hat{F})$ donde la función de distribución empírica $\hat{F}$ es un estimador de F. Usualmente el estimador *plug-in* es bueno si solamente la información existente sobre F viene de la muestra, caso contrario si existe información adicional por otro medio, el estimador *plug-in* no es eficiente.

3.6 Principio Básico del Bootstrap

El método *bootstrap* conceptualmente es uno de los procedimientos estadísticos computacionales más simples. Su base es la utilización extrema del principio *plug-in*.

El método de *bootstrap* en su situación general tal como describen *Efron y Tibshirani* (1986), es como sigue:

Sea $X=\{X_1, X_2, \ldots, X_n\}$ un conjunto de datos no necesariamente idéntica e independientemente distribuidos, generados por el modelo estadístico P, y sea el estadístico $T(X)$ cuya distribución $\mathbf{L}(T;P)$ se desea estimar. El método *bootstrap* propone como estimador de $\mathbf{L}(T;P)$ la distribución $\mathbf{L}^*(T^*;\hat{P})$ del estadístico $T^*=T(X^*)$ donde X^* es el conjunto de datos generado por el modelo $\hat{P}$. Notemos que si $\hat{P}=P$, entonces las distribuciones $\mathbf{L}(T;P)$ y $\mathbf{L}^*(T^*;\hat{P})$ coinciden, en algunos casos coinciden aun cuando $\hat{P} \neq P$.

De manera que si tenemos un buen estimador de P, es lógico suponer que $\mathbf{L}^*(T^*;\hat{P})$ se aproximará a $\mathbf{L}(T;P)$; además se prueba que:

$$\rho_\infty(\mathbf{L}(T;P), \mathbf{L}^*(T^*;\hat{P})) \xrightarrow{P} 0 \qquad \textbf{(3.5)}$$

Donde $\xrightarrow{P}$ representa la convergencia en probabilidad y ρ_∞ representa la distancia del supremo; así como se prueba $\rho_r(\mathbf{L}(T;P), \mathbf{L}^*(T^*;\hat{P})) \xrightarrow{P} 0$, donde ρ_r representa la distancia de *Mallows*, definida en el apéndice.

Este principio del *bootstrap* es más evidente en el caso donde $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidos (iid).

3.7 El Bootstrap IID

Un esquema de remuestreo no paramétrico en el contexto de los datos independientes e idénticamente distribuidos, es introducido como el *bootstrap* IID [Efron, 1979], también es llamado *navie bootstrap* ó *bootstrap ordinario*.

La formulación del método *bootstrap* IID, asume que $X_1, X_2, \ldots$ es una secuencia de variables aleatorias iid con distribución de probabilidad F desconocida. Supongamos que se genera la data $\chi_n = \{X_1, \ldots, X_n\}$ de F, por otro lado sea $T_n = t_n(\chi_n, F)$ la variable aleatoria de interés, con $n \geq 1$, denotemos a G_n como la distribución muestral de T_n, debemos encontrar una aproximación exacta de la distribución desconocida de T_n o alguna característica poblacional, como el error estándar de T_n, para ello el *bootstrap* provee una solución efectiva en la dirección del problema, sin imponer asumpciones sobre el modelo F.

Dada la muestra $\chi_n = \{X_1, \ldots, X_n\}$, seleccionamos una muestra aleatoria simple $\chi_m^* = \{X_1^*, \ldots, X_m^*\}$ de tamaño *m* con reemplazo de χ_n, llamada ***muestra bootstrap***. Así condicionado sobre χ_n, $X_1^*, \ldots, X_m^*$ son variables aleatorias iid, con

$$P_*(X_k^* = X_i) = \frac{1}{n}, \quad 1 \leq i \leq n, \text{ Para cada } k = 1, \ldots, m \qquad \textbf{(3.6)}$$

Donde P_* es la probabilidad condicional dado χ_n, por lo tanto la distribución común de los X_k^*, esta dada por la distribución empírica

$$F_n(\cdot) = n^{-1} \sum_{i=1}^{n} \mathrm{I}(X_i \leq \cdot), \qquad \textbf{(3.7)}$$

Donde I es la función indicadora; luego se define la versión *bootstrap* $T^*_{m,n}$ de T_n, mediante el reemplazo de χ_n con χ^*_m y F con F_n, como:

$$T^*_{m,n} = t_m(\chi^*_m; F_n) \qquad \textbf{(3.8)}$$

Denotemos también a $\hat{G}_{m,n}$ la distribución condicional de $T^*_{m,n}$, dado χ_n, por el principio del *bootstrap* $\hat{G}_{m,n}$ es un estimador de la distribución muestral desconocida G_n de T_n, de esta forma se puede estimar una funcional $\varphi(G_n)$ de la distribución muestral de T_n, por su correspondiente estimador *bootstrap plug-in* $\varphi(\hat{G}_{m,n})$.

Una vez que las variables χ_n han sido observadas, la distribución empírica F_n se convierte en conocida, y de esta manera es posible, al menos teóricamente encontrar la distribución condicional $\hat{G}_{m,n}$ y el estimador *bootstrap* $\varphi(\hat{G}_{m,n})$, a la par $\hat{G}_{m,n}$ es aproximado mediante simulación *Monte Carlo*.

En la práctica usualmente $\chi_n = \{X_1, \ldots, X_n\}$ representa una muestra aleatoria extraída de F; así como también se escoge el tamaño de la remuestra $m = n$, pero hay algunos trabajos en los cuales se considera $m \neq n$, ó toma un valor menor que *n*. Usualmente además se considera para la funcional $\varphi(G_n)$ expresiones como: $\varphi(G_n) = \mathrm{Var}(T_n) = \int x^2 dG_n(x) - \left(\int x dG_n(x)\right)^2$, donde su estimador *bootstrap* está dado por $\varphi(\hat{G}_{m,n}) = \mathrm{Var}(T^*_{m,n} \mid \chi_n) = \int x^2 d\hat{G}_{m,n}(x) - \left(\int x d\hat{G}_{m,n}(x)\right)^2$, así como $\varphi_\alpha(G_n) = \text{Cuantila } \alpha - \text{ésimo}$ con estimador *bootstrap* $\varphi_\alpha(\hat{G}_{m,n})$, etc. Por otro lado cabe

mencionar que encontrar exactamente $\hat{G}_{m,n}$ puede ser una tarea laboriosa aun en una muestra de tamaño moderado, esto debido a que el número de distintos posibles valores de χ_m^* crece rápidamente bajo el *bootstrap* IID.

Teóricamente gran cantidad de trabajos se dedicaron a probar las propiedades de los estimadores *bootstrap* , en *Arcones y Giné* (1989) muestran bajo ciertas condiciones, sea $X_1, X_2, \ldots$ una secuencia de variables aleatorias iid con $EX_i^2 < \infty$, $m = n$, muestran que la distribución condicional $\hat{G}_{n,n}$ de $T_{n,n}^*$, generadas por el *bootstrap* iid provee una aproximación válida para distribución muestral G_n de T_n, esto es que $\hat{G}_{n,n} \approx G_n$ para $\sup_x |\hat{G}_{n,n}(x) - G_n(x)| = \sup_x |P_*(T_{n,n}^* \le x) - P(T_n \le x)| = o(1)$, cuando $n \to \infty$, donde $o(1)$ representa el orden de convergencia uno. Hasta este punto se muestra el *bootstrap* IID en su forma general, para estimar la distribución muestral de un estadístico, pero un uso muy importante que se le da, es poder obtener estimadores de la varianza del estadístico, consecuentemente su error estándar, esto debido a que existe muy pocos estimadores que tienen una forma explícita para su error estándar, únicamente lo que existen son resultados asintóticos que se basan en el teorema de *Mann-Withney*.

3.7.1 El Estimador Bootstrap del Error Estándar

Dada una muestra de una población con función de distribución F desconocida, se desea estimar un parámetro de interés $\theta = t(F)$ basada en la muestra x, para ello se encuentra el estadístico $\hat{\theta} = s(\mathrm{x})$, ahora es necesario conocer la precisión de $\hat{\theta}$ mediante $\mathrm{se}_F(\hat{\theta})$.

Considerando la muestra x se selecciona una muestra *bootstrap* o remuestra $\mathrm{x}^* = (x_1^*, \ldots, x_n^*)$ del mismo tamaño y escogida con reemplazo, de la población de n objetos $(x_1, \ldots, x_n)$, es:

$$\hat{F} \rightarrow (x_1^*, \ldots, x_n^*) \qquad \textbf{(3.9)}$$

Donde $\hat{F}$ es la distribución empírica de x. Sea $\hat{\theta}^* = s(\mathrm{x}^*)$ el estimador del parámetro θ^* de la muestra *bootstrap* o resultado de aplicar la misma función $s(\cdot)$ a x^*, entonces el estimador *bootstrap* ideal del error estándar es:

$$\mathrm{se}_{\hat{F}}(\hat{\theta}^*). \qquad \textbf{(3.10)}$$

Este estimador *bootstrap* es propiamente un estimado *plug-in* por naturaleza, que usa la distribución empírica $\hat{F}$. En *Shao y Tu* (1995) se muestra algunas propiedades del estimador $\mathrm{se}_{\hat{F}}(\hat{\theta}^*)$, como resultados de consistencia, bajo ciertas restricciones.

Un punto importante es el cálculo de los estimadores *bootstrap*, ya que se tiene $\binom{\mathrm{n+m-1}}{\mathrm{m-1}}$ muestras distintas en la versión *bootstrap*, donde n es el tamaño de muestra original y m es el tamaño de muestra *bootstrap*, no se tiene fórmulas explícitas sobre los estimadores *bootstrap*, así el problema se aborda mediante, soluciones analíticas o aproximaciones como el método delta, el método de punto silla [Davison y Hinkley, 1997].

3.7.2 Algoritmo Bootstrap para Estimar Errores Estándar

Para obtener estimadores *bootstrap* de una muestra $\mathrm{x} = (x_1, \ldots, x_n)$ proveniente de la distribución empírica $\hat{F}$, se sigue el procedimiento *Monte Carlo:*

1. Seleccionar B muestras *bootstrap* independientes $\mathrm{x}^{*1}, \mathrm{x}^{*2}, \ldots, \mathrm{x}^{*B}$, cada una con *n* valores extraídos con reemplazo de la muestra original x.
2. Calcular el estadístico en cada muestra *bootstrap*, esto es $\hat{\theta}^*(b) = s(\mathrm{x}^{*b})$ para $b = 1, 2, \ldots, B$.
3. Estimar el error estándar $\mathrm{se}_F(\hat{\theta})$ por la desviación estándar de las repeticiones del estadístico, en las B muestras *bootstrap*, esto es

$$\widehat{\mathrm{se}}_B = \left\{ (B-1)^{-1} \sum_{b=1}^{B} (\hat{\theta}^*(b) - \overline{\hat{\theta}^*})^2 \right\}^{1/2} \qquad \textbf{(3.11)}$$

Donde $\overline{\hat{\theta}^*} = B^{-1} \sum_{b=1}^{B} \hat{\theta}^*(b)$. Cabe mencionar que la selección de B no es un problema sencillo, de acuerdo a la complejidad del problema uno puede realizar una elección prudente, existen algunas sugerencias como 25-200 [Efron y Tibshirani, 1986], para estimadores de momentos, y al menos 1000 remuestras para estimadores de distribuciones o cuantiles. Por otro lado se tiene que $\lim_{B \to \infty} \widehat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}}(\hat{\theta}^*)$, así el limite de $\widehat{\mathrm{se}}_B$ es el "estimador *bootstrap* ideal" de $\mathrm{se}_F(\hat{\theta})$; el hecho que $\widehat{\mathrm{se}}_B$ se aproxime a $\mathrm{se}_{\hat{F}}$ cuando B se aproxima a infinito, se dice la desviación estándar empírica se aproxima a la desviación estándar poblacional cuando el número de replicas crece. La población en este caso son todos los $\hat{\theta}^* = s(\mathrm{x}^*)$, donde $\hat{F} \to (x_1^*, x_2^*, \ldots, x_n^*) = \mathrm{x}^*$.

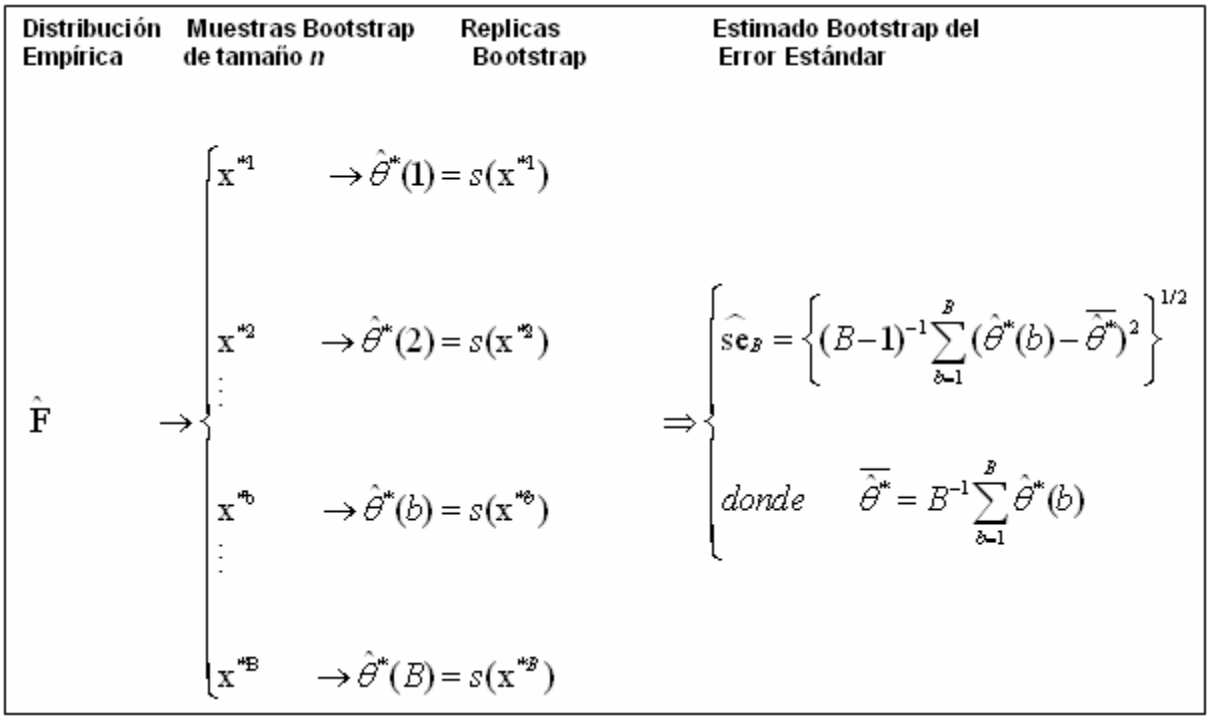


Figura 3.1: Algoritmo Bootstrap para Estimar el error Estándar de un Estadístico.

El estimador *bootstrap* ideal $\text{se}_{\hat{F}}(\hat{\theta}^*)$ y la aproximación $\widehat{\text{se}}_B$ algunas veces son llamados estimados *bootstrap* no paramétricos, porque están basados en $\hat{F}$, el cual es estimador no paramétrico del poblacional F.

3.8 Inadecuidad del Bootstrap IID para datos dependientes

El *bootstrap* IID es un método simple general que tiene su aplicación en problemas estadísticos, sin embargo la percepción general que el *bootstrap* es un método de antología, dando resultados automáticos exactos en todos los problemas, es erróneo. Una prueba de ello aparece en el trabajo de *Singh* (1981), en el cual se ve la primera confirmación teórica de superioridad del *bootstrap* IID, como también señala la inadecuidad para data dependiente.

Considerando el primer problema para datos dependientes; es que en la escena del *bootstrap* iid se impone independencia mutua sobre X_j, de esta forma asumiendo que su función de distribución conjunta es $F(x_1)\times F(x_2)\times\ldots\times F(x_n)$, así muestreando de esta distribución estimamos $\hat{F}(x_1^*)\times\hat{F}(x_2^*)\times\ldots\times\hat{F}(x_n^*)$, lo cual es incorrecto para X_j dependientes. Un segundo punto de afrontar este problema es de la forma siguiente. Supongamos que $\{X_t\}_{t\in\mathbb{Z}}$ es una secuencia de variables aleatorias **m-dependientes** con media μ y $E(X_t^2)<\infty$, denotemos con $\{X_n\}$ una muestra de tamaño *n* de la secuencia dada, luego sea $\sigma_m^2=\mathrm{Var}(X_t)+2\sum_{i=1}^{m}\mathrm{Cov}(X_t,X_{t+i})$ y $\bar{X}_n=n^{-1}\sum_{i=1}^{n}X_i$. Si $\sigma_m^2\in(0,\infty)$, luego por el teorema de límite central para variables **m-dependientes.**

$$\sqrt{n}\left(\bar{X}_n-\mu\right)\xrightarrow{d}N(0,\sigma_m^2) \qquad \textbf{(3.12)}$$

Ahora supongamos que deseamos estimar la distribución muestral de la variable aleatoria $T_n=\sqrt{n}\left(\bar{X}_n-\mu\right)$ usando el *bootstrap* IID. Asumiendo que el tamaño de las muestras *bootstrap* es igual al tamaño de muestra, esto es, de $\chi_n=\{X_1,\ldots,X_n\}$ son generadas un igual número de variables $X_1^*,\ldots,X_n^*$, entonces la versión *bootstrap* $T_{n,n}^*$ de T_n, esta dado por:

$$T_{n,n}^*=\sqrt{n}\left(\bar{X}_n^*-\mu\right) \qquad \textbf{(3.13)}$$

Donde $\bar{X}_n^*=n^{-1}\sum_{i=1}^{n}X_i^*$. La distribución condicional de $T_{n,n}^*$ bajo el *bootstrap* IID aun converge a la distribución normal, pero con una varianza errónea, esto es que, si bien

$$\sup_x|\,P_*(T_{n,n}^*\le x)-\Phi(x/\sigma)\,|=o(1),\ \text{cuando}\ \ n\to\infty, \qquad \textbf{(3.14)}$$

Donde $o(1)$ denota el orden de convergencia uno; el $\lim_{x\to\infty}[P_*(T^*_{n,n} \le x) - P(T_n \le x)] = [\Phi(x/\sigma) - \Phi(x/\sigma_\infty)] \neq 0$, si $\sum_{i=1}^{m} \mathrm{Cov}(X_t, X_{t+i}) \neq 0$ y $\sigma^2_\infty \neq 0$, para $x \neq 0$. Donde $P_*(T^*_{n,n} \le x) = \hat{G}_{n,n}(x)$ es la distribución muestral del estadístico $T^*_{n,n}$.

Así para todo $x \neq 0$, el estimador *bootstrap* IID $P_*(T^*_{n,n} \le x)$ de un parámetro de nivel-2 de $P(T_n \le x)$, tiene un error cuadrático medio que no tiende a cero en el límite y así, no es consistente, esto por $\hat{G}_{n,n}(x) \neq G_n(x)$. Así el método de *bootstrap* IID falla drásticamente para datos dependientes, que trata de remuestrear los X_i's de la data χ_n, ignorando completamente la estructura de dependencia en ella, así como también falla la explicación de la covarianza retrasada en relación a la varianza asintótica. Finalmente una prueba rápida pero eficaz de la ineficiencia del *bootstrap* IID para datos dependientes, sea la secuencia dependiente $\{X_t\}_{t\in\mathbb{Z}}$, con $\mathrm{Var}(X_t)=\sigma^2$ y autocorrelaciones $\rho_h = \mathrm{corr}(X_t X_{t+h})$ para $h = 1, 2, \ldots$; en la estimación del la varianza de $\bar{X}_t$

$$Var(\bar{X}_t) = \frac{\sigma^2}{n} \sum_{h=-(n-1)}^{n-1} \left(1 - \frac{|h|}{n}\right) \rho_h \qquad \textbf{(3.15)}$$

La sumatoria difiere considerablemente de uno, así el estimador *bootstrap* de la varianza será erróneo e inconsistente. Así siguiendo estos resultados del *bootstrap* IID, se requiere considerar nuevas extensiones para datos dependientes, que puedan reproducir su estructura de dependencia o puedan tener un funcionamiento alternativo para tratar el problema de dependencia. Es así que surgen nuevas alternativas dentro del tratado de datos dependientes, datos longitudinales, datos espaciales, los cuales presentan dicha particularidad; como ya es

sabido el presente se orienta dentro de la estructura de datos dependientes, más propiamente dichos datos extraídos a lo largo del tiempo, usualmente llamados datos temporales o series de tiempo, que presentan en su estructura distintos tipos de dependencia.

Los métodos de remuestreo en series temporales son divididos en dos grandes ramas, bajo el criterio si utilizan un modelo para la estructura de dependencia, ó no [Li y Maddala, 1996].

1. El enfoque basado en modelo (BM).
2. El enfoque no basado en modelo (NBM).

El primero aplica un método de remuestreo a los residuos obtenidos a partir de ajustar los datos dependientes un modelo adecuado, sea paramétrico o no. El segundo usualmente aplica un método de remuestreo a bloques de observaciones consecutivas de los datos dependientes. Dentro de los dos enfoques existen diversas metodologías y técnicas para la utilización de un método adecuado de remuestreo y pautas importantes para el desarrollo de estos. Inicialmente se desarrollaron los métodos basados en modelos, debido a la alta analogía presentada con los métodos de remuestreo para regresión lineal con errores iid, donde se remuestrean los residuos; así siguiendo este último hecho, en datos dependientes inicialmente se ajusta un modelo adecuado y luego se remuestrean las innovaciones, siguiendo premisas adecuadas; el otro enfoque que surgió a la par con el anterior, y tiene una base teórica importante, el cual es, que las funciones de autocovarianza de los bloques de observaciones sean lo más parecidas posibles, obviamente teniendo premisas importantes sobre la data dependiente. Este método no basado en modelo, de alguna forma no reproduce la estructura de dependencia de la data a pesar de manejar ciertos parámetros como la amplitud de cada bloque de observaciones consecutivas, o considerar si los bloques son solapados o no,

adicionalmente la dependencia entre diferentes bloques no es tomada en cuenta [Künsch,1989], al margen de todo ello, como un objetivo de las series temporales son las predicciones, este método no permite reproducir la estructura de dependencia del modelo, por esta razón los predictores lineales se ven afectados grandemente; es así que el enfoque basado en modelo será nuestra principal dirección con referencia al trabajo.

Dos principales características se presentan para la utilización del enfoque basado en modelo una de las cuales es la elección de un modelo paramétrico o no, para el comportamiento de la serie de tiempo, para ello existen diversas metodologías para la elección de un modelo adecuado para datos temporales así como una gran familia de modelos lineales y no lineales para ajustar la serie. Finalmente para completar la metodología es importante seleccionar un método de remuestreo adecuado que nos permita reproducir la estructura de dependencia. A continuación se presenta el *bootstrap* como una alternativa para el remuestreo y modelos autorregresivos para el modelo.

3.9 Bootstrap Basado en Modelo

En el presente enfoque se considera el método de *bootstrap* para algunos modelos de series de tiempo, como los procesos autoregresivos que son guiados por variables aleatorias iid a través de una ecuación estructural; para modelar esto, es posible adaptar las ideas básicas del *bootstrapping* en un modelo de regresión lineal con errores variables iid [Freedman, 1981], y el esquema de remuestreo IID, que remuestrea solo un valor en un tiempo [Efron, 1979].

3.9.1 Bootstrap en Innovaciones IID

De manera general para tratar este enfoque se presenta el *bootstrap* que actúa en las innovaciones de un modelo propuesto, y se formula la metodología para obtener estimadores *bootstrap* deseados, *Efron y Tibshirani* (1993), esto es. Sea $\{X_t, t \in \mathrm{Z}\}$ una secuencia de variables aleatorias que satisfacen la ecuación:

$$X_t = h\left(X_{t-1}, \ldots, X_{t-p}; \beta\right) + \varepsilon_t \qquad \textbf{(3.16)}$$

Donde β es un vector de parámetros $q \times 1$, $h : \mathbb{R}^{p+q} \to \mathbb{R}$ es una función medible borel conocida, con $n > p$, p es el orden autorregresivo, y $\{\varepsilon_t\}$ es una secuencia de variables aleatorias iid con distribución común F, que son independientes con las variables aleatorias $X_1, \ldots, X_p$. Para la identificación de un modelo (3.16) se asume que $E(\varepsilon_t) = 0$. Como el proceso $\{X_t, t \in \mathrm{Z}\}$ es conducido por las innovaciones ε_i que son iid, el método de *bootstrap* IID puede ser extendido para el modelo dependiente (3.16). Sin perder la generalidad supongamos $\chi_n = \{X_1, \ldots, X_n\}$ una realización del proceso y se desea aproximar la distribución muestral de la variable aleatoria $T_n = t_n\left(\chi_n; F, \beta\right)$. Sea $\hat{\beta}_n$ el estimador mínimo cuadrático de β basado en χ_n, así se define los residuales $\hat{\varepsilon}_i = X_i - h\left(X_{i-1}, \ldots, X_{i-p}; \hat{\beta}_n\right)$, con $p < i \le n$, luego se centran los residuales y se define $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \overline{\varepsilon}_n$, con $p < i \le n$, donde $\overline{\varepsilon}_n \equiv (n-p)^{-1} \sum_{i=1}^{n-p} \hat{\varepsilon}_{i+p} \neq 0$; si no se centralizan los residuos los resultados de la aproximación *bootstrap* usualmente tienen un sesgo aleatorio que no se desaparece en el límite y da una aproximación inútil *Lahiri* (2003).

Luego se extrae una muestra aleatoria simple con remplazamiento $\varepsilon^*_{p+1},\ldots,\varepsilon^*_m$ de tamaño $(m-p)$ de $\{\tilde{\varepsilon}_i : p < i \le n\}$ y definimos el *bootstrap* de las pseudos-observaciones, usando la estructura del modelo (3.16), como:

$$X^*_i = X_i \quad \text{para } i = 1,\ldots,p \quad \text{y} \tag{3.17}$$

$$X^*_i = h\left(X^*_{i-1},\ldots,X^*_{i-p};\hat{\beta}_n\right) + \varepsilon^*_i \quad \text{para } p < i \le m \tag{3.18}$$

Donde ε^*_i para $p < i \le m$ son iid y $E_*(\varepsilon^*_t) = 0$. La versión *bootstrap* de la variable aleatoria $T_n = t_n(\chi_n; F, \beta)$ esta definido como $T^*_{m,n} = t_m\left(\chi^*_m; F_n, \hat{\beta}_n\right)$, donde $\chi^*_m = \{X^*_1,\ldots,X^*_m\}$ y F_n denota la distribución empírica de los residuales centrados $\tilde{\varepsilon}_i : p < i \le n$. La distribución muestral de T_n, es aproximado por la distribución condicional de $T^*_{m,n}$ dado χ_n. Para diferentes esquemas del modelo (3.16), existen diferentes versiones de remuestreo que fueron propuestos. Un caso especial es considerar un modelo autoregresivo de orden p dado por $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t$, donde $\beta = (\phi_1, \phi_2 \ldots, \phi_p)$ es el vector de parámetros autorregresivos, y $\{\varepsilon_t\}$ es una secuencia iid, que satisface los requerimientos de (3.16).

3.9.2 Bootstrapping en Procesos Autoregresivos Estacionarios

La estructura de dependencia de una serie de tiempo es usualmente representada por el pasado de la serie, es así que los procesos autoregresivos tiene un gran uso en la teoría de predicción, y con relación al *bootstrap* en innovaciones se presenta el *bootstrapping* en procesos autoregresivos *Efron y Tibshirani* (1993) como:

Sea $\{X_t, t \in \mathrm{Z}\}$ un proceso autoregresivo estacionario de orden p, $AR(p)$, que satisface la ecuación lineal

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t \qquad \textbf{(3.19)}$$

Donde $p \in \mathbb{N}$, $\phi_1, \phi_2 \ldots, \phi_p$ son los parámetros autoregresivos y $\{\varepsilon_t\}$ es una secuencia de variables aleatorias iid con media cero y distribución común F; en adición usualmente se asume que los parámetros autoregresivos son tales que $\phi(z) \equiv 1 - \sum_{j=1}^{p} \phi_j z^j \neq 0$, para todo $z \in \mathbb{C}$, con $|z| \leq 1$, [Brockwell y Davis, 2002], y que bajo esta propiedad el proceso $\{X_t\} \sim AR(p)$ admite una representación media móvil de orden infinito, esto es

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \qquad \textbf{(3.20)}$$

Donde $\{\psi_j\}_{j=0}^{\infty}$ son constantes, determinadas por la expansión de la serie de potencia de la función $\psi(z) \equiv [\phi(z)]^{-1}$ dado por:

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j, \quad |z| \leq 1 \qquad \textbf{(3.21)}$$

Aun las variables aleatorias X_i's bajo el modelo *AR(p)* son dependientes, podemos usar la estructura del modelo para generar una aproximación válida del *bootstrap*; la idea básica es considerar los residuales de ajustar el modelo, los cuales son aproximadamente independientes, luego remuestrear los residuales. Supongamos que $X_1, \ldots, X_n$ es una realización o serie de tiempo del proceso $\{X_t\} \sim AR(p)$; sea $\hat{\phi}_{1n}, \hat{\phi}_{2n}, \ldots, \hat{\phi}_{pn}$ los estimadores

mínimo cuadráticos de $\phi_1, \phi_2 \ldots, \phi_p$, basados en $X_1, \ldots, X_n$. Así $\hat{\phi}_{1n}, \hat{\phi}_{2n}, \ldots, \hat{\phi}_{pn}$ están dados por la relación

$$(\hat{\phi}_{1n}, \hat{\phi}_{2n}, \ldots, \hat{\phi}_{pn})' = (V_n' V_n)^{-1} V_n' (X_{p+1}, \ldots, X_n)' \quad \textbf{(3.22)}$$

Donde V_n es una matriz de orden $(n-p) \times p$, con la i-esima fila $(X_{i+p+1}, \ldots, X_i)$, $i = 1, \ldots n-p$. Sea $\hat{\varepsilon}_i = X_i - \hat{\phi}_{1n} X_{i-1} - \hat{\phi}_{2n} X_{i-2} - \ldots - \hat{\phi}_{pn} X_{i-p}$, $i = p+1, \ldots, n$ los residuales. Por (3.19) los residuales se pueden expresar como

$$\hat{\varepsilon}_i = \varepsilon_i - \sum_{j=1}^{p} (\hat{\phi}_{jp} - \phi_j) X_{i-j}, \quad p+1 \le i \le n \quad \textbf{(3.23)}$$

Cuando $\hat{\phi}_{jn} \xrightarrow{p} \phi_j$ para $j = 1, \ldots, p$, el segundo término de (3.23) es pequeño para valores grandes de *n*, de esta manera los residuales son aproximadamente independientes; así podemos remuestrear un solo valor en el tiempo, los residuales, con el *bootstrap* IID, para definir la versión *bootstrap* de una variable aleatoria $T_n = t_n(X_1, \ldots, X_n; \phi_1, \ldots, \phi_p, F)$. Sin embargo para generar una aproximación válida se necesita centrar los residuales, y luego el proceso de remuestreo a la colección de los residuales centrados, definidos por

$$\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \bar{\varepsilon}_n, \quad i = p+1, \ldots, n \quad \textbf{(3.24)}$$

Donde $\bar{\varepsilon}_n = (n-p)^{-1} \sum_{i=p+1}^{n} \hat{\varepsilon}_i$, luego generamos los errores *bootstrap* ε_i^*, $i \in \mathbb{Z}$, mediante un muestreo aleatorio simple con remplazamiento de $\{\tilde{\varepsilon}_{p+1}, \ldots, \tilde{\varepsilon}_n\}$, así las variables aleatorias ε_i^*, $i \in \mathbb{Z}$ son condicionalmente iid dado $X_1, \ldots, X_n$, con distribución común

$$P_*(\varepsilon_t^* = \tilde{\varepsilon}_i) = \frac{1}{n-p} \quad \textbf{(3.25)}$$

Por (3.24) y (3.25) $E_*(\varepsilon_t^*) = (n-p)^{-1}\sum_{i=p+1}^{n}\tilde{\varepsilon}_i = 0$, así los errores *bootstrap* ε_i^* satisfacen el análogo $E(\varepsilon_t) = 0$. Luego se define la versión *bootstrap* de la ecuación (3.19)**.**

$$X_t^* = \hat{\phi}_{1n}X_{t-1}^* + \hat{\phi}_{2n}X_{t-2}^* + \ldots + \hat{\phi}_{pn}X_{t-p}^* + \varepsilon_t^* \qquad \textbf{(3.26)}$$

Sea $\{X_i^*\}_{i\in\mathbb{Z}}$ una solución estacionaria de (3.26). Si $\hat{\phi}_{jn} \xrightarrow{p} \phi_j$, cuando $n \to \infty$, para $j = 1,\ldots, p$, entonces tal solución existe sobre el conjunto de los X_i que tiene probabilidad cerca de uno, cuando *n* es grande. En la practica para usar la recursion (3.26) para $i \geq p+1$ generamos las observaciones *bootstrap,* fijando los *p* valores iniciales iguales a $X_1,\ldots, X_p$, o ceros. Cuando el polinomio $\hat{\phi}(z) \equiv 1 - \sum_{j=1}^{p}\hat{\phi}_{jn}z^j \neq 0$ no desaparece en la región $\{|z| \leq 1\}$, los coeficientes de los *p* valores iniciales se extinguen geométricamente rápido, y así tiene un insignificante efecto a largo plazo. La versión *bootstrap* autorregresiva de una variable aleatoria $T_n = t_n(X_1,\ldots, X_n; \phi_1,\ldots,\phi_p, F)$ basado en un remuestreo de tamaño $m > p$ es

$$T_{m,n}^* = t_m(X_1^*,\ldots, X_n^*; \hat{\phi}_{1n},\ldots,\hat{\phi}_{pn}, \hat{F}_n) \qquad \textbf{(3.27)}$$

Donde $\hat{F}_n$ es la distribución empírica de los residuales centrados $\tilde{\varepsilon}_i$, $i = p+1,\ldots, n$, y típicamente el tamaño de las remuestras *m* son iguales al tamaño de muestra original *n.*

4 SIEVE BOOTSTRAP

4.1 Introducción

En el enfoque de los métodos de remuestreo basados en modelo para datos dependientes, específicamente en el contexto de las series de tiempo estacionarias, se presenta el estudio de un método de *bootstrap* el cual esta basado en el método de *sieves* [Grenander, 1981]; comúnmente llamado *sieve bootstrap,* que tiene como idea básica ajustar inicialmente un modelo paramétrico y luego remuestrear los residuales. Pero en vez de considerar un modelo fijo finito dimensional, considera aproximar un modelo no paramétrico infinito dimensional, mediante una secuencia de modelos parametricos finito dimensionales. Implícitamente, esta aproximación es usualmente usada cuando se escoge un modelo adaptativamente, mediante un criterio de elección, como el *Akaike Information Criterion* (AIC) que considera un modelo prefijado. De esta manera se aproxima un verdadero esencial proceso estacionario, mediante un modelo autorregresivo de orden *p,* donde $p = p(n)$ es una función del tamaño de muestra *n*, con determinadas propiedades. Luego se estima un modelo $AR\big(p(n)\big)$, y se genera una muestra *bootstrap* mediante el remuestreo de los residuales; así el *sieve bootstrap* goza de propiedades no paramétricas buenas, siendo un modelo libre dentro de una clase de procesos lineales.

Como se mostró anteriormente el *sieve bootstrap* en su forma general se presenta de la siguiente manera

Sea $\{X_t\}_{t\in\mathbb{Z}}$ una serie de tiempo estacionaria y sea $T_n = t_n(X_1,\ldots,X_n)$ un estimador de un parámetro de nivel-1 de interés $\theta = \theta(P)$ donde P denota la distribución conjunta de probabilidad (desconocida) de $\{X_t\}_{t\in\mathbb{Z}}$, entonces la distribución muestral de T_n esta dado por

$$G_n(B) = P(T_n \in B) \qquad \textbf{(4.1)}$$

Donde B es un conjunto de *Borel* en $\mathbb{R}$, como es sabido el *bootstrap* estima parámetros de nivel-2 como $G_n(B)$, $Var(T_n)$, etc. Como los X_i's son dependientes la estimación de un parámetro de nivel-2 $G_n(B)$ puede ser pensado como un procedimiento de dos pasos, el primero es aproximar P mediante una simple distribución de probabilidad $\tilde{P}_n$, el siguiente paso es estimar $\tilde{P}_n$ usando la data $\{X_1,\ldots,X_n\}$. La idea del *sieve bootstrap* es escoger $\{\tilde{P}_n\}_{n\geq 1}$ para hacer una aproximación fina de P. esto es que $\{\tilde{P}_n\}_{n\geq 1}$ es una secuencia de medidas de probabilidad sobre $\left(\mathbb{R}^\infty, \mathbf{B}(\mathbb{R}^\infty)\right)$, tal que para cada n, $\tilde{P}_{n+1}$ es una aproximación fina para P que $\tilde{P}_n$, y $\tilde{P}_n$ converge a P cuando $n \to \infty$.

4.2 Definición del Sieve Bootstrap

Consideremos un proceso estacionario $\{X_t\}_{t\in\mathbb{Z}}$ de valores reales, con esperanza $E(X_t) = \mu_X$, si $\{X_t\}_{t\in\mathbb{Z}}$ es puramente no determinístico, por el teorema de *Wold* podemos escribir $\{X_t - \mu_X\}_{t\in\mathbb{Z}}$, como un proceso de media móvil de orden infinito

$$X_t - \mu_X = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \psi_0 = 1 \tag{4.2}$$

Donde $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ es una secuencia de variables no correlacionadas con $E(\varepsilon_t) = 0$ y $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, además de requerir que el proceso (4.2) sea invertible, el cual nos limita a una pequeña clase de procesos estacionarios. Bajo la adicional suposición de invertibilidad [Brockwell y Davis, 1987], se puede representar $\{X_t\}_{t\in\mathbb{Z}}$ como un proceso autorregresivo de orden infinito

$$\sum_{j=0}^{\infty} \phi_j (X_{t-j} - \mu_X) = \varepsilon_t, \quad \phi_0 = 1 \tag{4.3}$$

Con $\sum_{j=0}^{\infty} \phi_j^2 < \infty$; así la representación (4.3) se puede considerar como una aproximación autorregresiva como un *sieve* para el proceso estocástico $\{X_t\}_{t\in\mathbb{Z}}$. Mediante (4.2) también se puede usar la aproximación medias móviles, pero la aproximación autorregresiva el cual, como un método lineal, es más popular, rápido y bien conocido como una técnica satisfactoria en diferentes situaciones.

La definición formal del *sieve bootstrap* es como sigue, sea $X_1, X_2, \ldots, X_n$ una muestra del proceso $\{X_t\}_{t\in\mathbb{Z}}$, se ajusta un proceso autorregresivo, con un orden que crece $p(n)$ cuando el tamaño de muestra n crece. Sea $p = p(n) \to \infty, \quad n \to \infty$ con $p(n) = o(n)$ donde $o(\cdot)$ representa el orden de convergencia de p, esto es que $p(n) = o(n)$ si $p(n)/n \to 0$ en el límite cuando $n \to \infty$, esto implica que si bien suponemos que el orden del modelo puede crecer con *n*, ese crecimiento no puede ser muy rápido, al menos no tan rápido como el de *n*.

Luego estimamos los coeficientes del modelo $\hat{\phi}_{1,n}, \hat{\phi}_{2,n}, \ldots, \hat{\phi}_{p,n}$, correspondientes al modelo (4.3) usualmente mediante los estimadores *Yule Walker* [Brockwell y Davis, 1987], inicialmente se suele sustraer la media muestral $\bar{X}$, así los residuales se computan

$$\hat{\varepsilon}_{t,n} = \sum_{j=0}^{p(n)} \hat{\phi}_{j,n}(X_{t-j} - \bar{X}), \quad \hat{\phi}_{0,n} = 1, \quad \text{con } t = p+1, \ldots, n \tag{4.4}$$

Luego construimos el remuestreo basado en esta aproximación autorregresiva. Previamente centramos los residuales mediante

$$\tilde{\varepsilon}_{t,n} = \hat{\varepsilon}_{t,n} - \left[(n-p)^{-1} \sum_{t=p+1}^{n} \hat{\varepsilon}_{t,n} \right], \quad t = p+1, \ldots, n \tag{4.5}$$

Y denotamos la función de distribución empírica de $\{\tilde{\varepsilon}_{t,n}\}_{t=p+1}^{n}$, mediante

$$\hat{F}_{\varepsilon,n}(\cdot) = (n-p)^{-1} \sum_{t=p+1}^{n} I_{[\tilde{\varepsilon}_{t,n} \leq \cdot]} \tag{4.6}$$

Donde I, representa la función indicadora; luego se puede remuestrear para algún $t \in \mathbb{Z}$

$$\varepsilon_t^* \ \ i.i.d. \sim \hat{F}_{\varepsilon,n} \tag{4.7}$$

Luego se define $\{X_t^*\}_{t \in \mathbb{Z}}$, mediante la recursion

$$\sum_{j=0}^{p(n)} \hat{\phi}_{j,n}(X_{t-j}^* - \bar{X}) = \varepsilon_t^* \tag{4.8}$$

Consecuentemente se construye la muestra *sieve bootstrap* $X_1^*, X_2^*, \ldots, X_n^*$, escogiendo los valores iniciales iguales a $\bar{X}$, generando un proceso $AR\big(p(n)\big)$ como en (4.8) hasta que la estacionariedad se alcance y luego proyectar fuera los primeros valores generados. Esta

aproximación *bootstrap* induce una probabilidad condicional P^* dado la muestra $X_1, X_2, \ldots, X_n$.

Considerando algún estadístico $T_n = T_n(X_1, \ldots, X_n)$, donde T_n es una función medible de *n* observaciones, así se define el estadístico *Bootstrap* T_n^* mediante el principio *Plug-in*

$$T_n^* = T_n(X_1^*, \ldots, X_n^*) \qquad \textbf{(4.9)}$$

Este procedimiento *bootstrap* exhibe condiciones mejores y más simples en comparación con el punto de vista de *bootstrap* por bloques. No necesita pre-vectorizar las observaciones originales; también es tratado del *sieve* es mas simple en el caso de tener datos faltantes [Bühlmann, 1997] algunas de las suposiciones y propiedades están dadas en el apéndice.

4.2.1 Elección del orden *p*

El procedimiento de aproximación *sieve* para un proceso autorregresivo en conjunción con el procedimiento para seleccionar modelos con mínimo AIC con innovaciones gausianas, permiten un método de elección del orden *p* del proceso. Para la predicción de un modelo $AR(\infty)$ es óptimo usar el AIC, más aún si se trata de estimar la varianza de la media, el AIC realiza una elección buena del orden.

Para conseguir estimadores *sieve bootstrap* para estimadores lineales y no lineales, requieren de condiciones de regularidad para el orden $p = p(n)$, de un proceso autorregresivos aproximado el cual cubre todas las situaciones generales, así este orden se puede considerar con un parámetro de suavizamiento, luego el problema se direcciona en escoger un valor

óptimo para p , de tal forma que el procedimiento *sieve bootstrap* sea exacto; existen dos procedimientos claros para la identificación de *p.*

Si el proceso $\{X_t\}_{t\in\mathbb{Z}}$ es un $AR(\infty)$, entonces el criterio AIC encabeza una teoría para la elección de orden p , que asintóticamente es eficiente la elección de $\hat{p}_{AIC}$ para el orden óptimo $p_{opt}(n)$ para algún proceso $AR(\infty)$ [Shibata, 1980], por otra parte el criterio BIC tiene propiedades optimas cuando el orden de un proceso autorregresivo es de orden finito, así el AIC es óptimo para un modelo complejo y no de dimensión finita. El BIC recae en la idea de prescribir como un mecanismo gráfico. Para algún *p,* se ajusta un modelo autorregresivo, obteniendo los residuales y calculando la densidad espectral estimada basada en los residuales, así escogeríamos un *p* tal que este espectro estimado este cerca de una constante, este último método puede detectar autocorrelación pero no es posible distinguir entre no correlacionado e innovaciones independientes [Lahiri, 2003].

4.3 Intervalos de Predicción basados en Sieve Bootstrap

Como un objetivo principal en el análisis de series de tiempo es predecir valores futuros de la data observada, y más específicamente como calcular intervalos de predicción, de esta manera se propone un procedimiento no paraméntrico *sieve bootstrap* $AR(\infty)$, para la construcción de intervalos de predicción para una clase general de modelos lineales que incluyen estacionariedad e invertibilidad *ARMA*, mediante un extensivo estudio *Monte Carlo*.

Sea $\{X_t\}_{t\in\mathbb{Z}}$ un proceso estacionario con valor esperado $E(X_t)=\mu_X$ que admite la representación $AR(\infty)$, el estudio *Monte Carlo* es como sigue:

1. Dada una muestra $\{X_1,\ldots,X_n\}$, seleccionamos $p = p(n)$ de una aproximación autorregresiva mediante el criterio de AIC.
2. Obtenemos los estimadores *Yule Walker* de los coeficientes autorregresivos: $(\hat{\phi}_1,\hat{\phi}_2,\ldots,\hat{\phi}_p)^T$.
3. Calcular los residuales $\hat{\varepsilon}_t = \sum_{j=0}^{p} \hat{\phi}_j (X_{t-j} - \bar{X}), \quad \hat{\phi}_0 = 1, \quad \text{con } t = p+1,\ldots,n$.
4. Calcular la función de distribución empírica de los residuales centrados: $\hat{F}_{\tilde{\varepsilon}}(x) = (n-p)^{-1} \sum_{t=p+1}^{n} I_{[\tilde{\varepsilon}_t \le x]}$, donde $\tilde{\varepsilon}_t = \hat{\varepsilon}_t - \left[(n-p)^{-1} \sum_{t=p+1}^{n} \hat{\varepsilon}_t \right]$
5. Extraemos remuestras ε_t^* de observaciones i.i.d de $\hat{F}_{\tilde{\varepsilon}}(x)$.
6. Definimos X_t^* mediante la recursión: $\sum_{j=0}^{p} \hat{\phi}_j (X_{t-j}^* - \bar{X}) = \varepsilon_t^*$, donde las p observaciones iniciales son iguales a $\bar{X}$.

Antes de este paso el esquema de remuestreo coincide con el *sieve bootstrap*, y es válido para la versión *bootstrap* de algunos estadísticos definidos como funcionales de una función de distribución *m*-dimensional. Pero esto no es eficaz para la predicción dado que no reproduce la distribución condicional de X_{T+h} dado la data observada. De este modo se debe fijar las últimas *p* observaciones [Cao, 1997], se puede obtener remuestras de los valores futuros X_{T+h}^* dado $X_{T-p+1}^* = X_{T-p+1},\ldots,X_T^* = X_T$

7. Calcular los estimadores $(\hat{\phi}_1^*,\hat{\phi}_2^*,\ldots,\hat{\phi}_p^*)^T$, análogo a 2.

8. Calcular las futuras observaciones *bootstrap* mediante:

$$X^*_{T+h} - \bar{X} = -\sum_{j=1}^{p} \hat{\phi}^*_j (X^*_{T+h-j} - \bar{X}) + \varepsilon^*_t$$

Donde $h > 0$ y $X^*_t = X_t$, para $t \leq T$.

Finalmente la distribución *bootstrap* $F^*_{X_{T+h}}$ de X^*_{T+h}, se usa para aproximar la distribución desconocida de X_{T+h}, dado la muestra observada. Así el intervalo de predicción de $(1-\alpha)\%$ para X_{T+h} esta dado por:

$$\left[Q^*(\alpha/2), Q^*(1-\alpha/2) \right] \qquad \textbf{(4.10)}$$

Donde $Q^*(\cdot)$ son las cuantilas de la distribución *bootstrap* estimada de las predicciones X_{T+h} algunas propiedades básicas están dadas en el apéndice.

Finalmente dado el estudio de series de tiempo climatológicas en la presente tesis; por lo general las series climatológicas y meteorológicas, presentan un comportamiento bastante particular; esto es que en su estructura es bien definida la componente estacional. Por tal razón se presenta la construcción de intervalos de predicción bajo el método de *sieve bootstrap*, presentando algunos cambios con referencia a lo expuesto anteriormente.

El procedimiento *Box Jenkins* considera un modelo $SARIMA(p,d,q)(P,D,Q)_S$, definido en (2.24). En tal caso si los órdenes con conocidos, los parámetros estimados asumiendo normalidad en los errores, entonces un intervalo de predicción es definido como en (2.42), este intervalo de confianza es afectado por la verdadera distribución de los errores, ya que al ser asimétrico los errores, la metodología *Box Jenkins* tiene un funcionamiento sesgado. Por otro lado el intervalo de predicción (2.42), no incorpora la incertidumbre de la estimación de

los parámetros, siendo esto importante cunando se trabaja con un tamaño de muestra pequeño, y cunado los errores no son gaussianos. Para este problema si los órdenes regular y estacional son conocidos, *Alonso* (2001) muestra que el $AR(\infty)$ -*sieve bootstrap* estacional, provee intervalos de predicción consistentes, con propiedades destacables no-paramétricas como, el de ser un modelo libre ya que no requiere normalidad en los errores, más aun trabaja con una distribución empírica para estos.

El *sieve bootstrap* para un proceso lineal que admite la representación $AR(\infty)$ estacional como:

$$\sum_{j=0}^{\infty}\phi_j B^j \sum_{j=0}^{\infty}\Phi_j B^{sj} Y_t = \varepsilon_t \qquad \textbf{(4.11)}$$

Donde $Y_t = (1-B)^d (1-B^s)^D X_t$, y los coeficientes $\{\phi_j\}_{j=0}^{\infty}$ y $\{\phi_j\}_{j=0}^{\infty}$, satisfacen para algún $r>2$ $\sum_{j=0}^{\infty} j^r |\phi_j| < \infty$ y $\sum_{j=0}^{\infty} j^r |\Phi_j| < \infty$, esto es cumplido por el modelo (4.11), desde que $\{\phi_j\}_{j=0}^{\infty}$ y $\{\phi_j\}_{j=0}^{\infty}$ tienen un decaimiento exponencial. A continuación se presenta el estudio *Monte Carlo* para la construcción de intervalos de predicción sieve bootstrap para (4.11).

Sea $\{X_t\}_{t\in\mathbb{Z}}$ un proceso que admite representación $AR(\infty)$ estacional como en (4.11), el método procede como sigue:

1. Dado una muestra $\{X_1, X_2, \ldots, X_T\}$, obtenemos la serie diferenciada $Y_t = (1-B)^d (1-B^s)^D X_t$, con $t = d + sD + 1, d + sD + 2, \ldots, T$.
2. Dado $\{Y_{d+sD+1}, Y_{d+sD+2}, \ldots, Y_T\}$ seleccionamos los ordenes $p = p(T)$ y $P = P(T)$ regular y estacional autorregresivo, mediante la aproximación del criterio

$AICC = -T\ln\sigma^2 + \frac{2(p+P+1)T}{T-p-P-2}$, con $0 \le p \le p_{\max}$ y $0 \le P \le P_{\max}$, el criterio de AICC es la versión del AIC corregido [Hurvich and Tsai, 1989]; esta corrección contrarresta el sobre ajuste natural de AIC.

3. Utilizamos algún método de estimación para obtener $\hat{\boldsymbol{\phi}}_{\boldsymbol{p}} = (\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p)'$ y $\hat{\boldsymbol{\Phi}}_{\boldsymbol{P}} = (\hat{\Phi}_1, \hat{\Phi}_2, \ldots, \hat{\Phi}_P)'$, supongamos escogemos el método de mínimos cuadrados, así $\hat{\boldsymbol{\phi}}_{\boldsymbol{p}}$ y $\hat{\boldsymbol{\Phi}}_{\boldsymbol{P}}$ son soluciones de: $\min_{\boldsymbol{\phi}_p, \boldsymbol{\Phi}_P} \left\{ \sum_{t=d+s(D+P)+p+1}^{T} \left(\sum_{j=0}^{p} \phi_j B^j \sum_{j=0}^{P} \Phi_j B^{sj} Y_t \right)^2 \right\}$

4. Calculando los residuales:

$$\hat{\varepsilon}_t = \sum_{j=0}^{p} \hat{\phi}_j B^j \sum_{j=0}^{P} \hat{\Phi}_j B^{sj} Y_t \; ; \quad \hat{\phi}_0 = 1, \hat{\Phi}_0 = 1, t \in \{d + s(D+P) + p + 1, ..., T\}.$$

5. Definir la función de distribución empírica de los residuales centrados:

$$\hat{F}_{\tilde{\varepsilon}}(x) = \frac{\sum_{t=d+s(D+P)+p+1}^{T} I\{\tilde{\varepsilon}_t \le x\}}{T - (d + s(D+P) + p)}, \text{ donde } \tilde{\varepsilon}_t = \hat{\varepsilon}_t - \hat{\varepsilon}^{(.)} \text{ y } \hat{\varepsilon}^{(.)} = \frac{\sum_{t=d+s(D+P)+p+1}^{T} \hat{\varepsilon}_t}{T - (d + s(D+P) + p)}$$

6. Extraemos remuestras ε_t^* i.i.d. observaciones de $\hat{F}_{\tilde{\varepsilon}}(x)$.

7. Definimos X_t^* mediante la recursion: $\sum_{j=0}^{p} \hat{\phi}_j B^j \sum_{j=0}^{P} \hat{\Phi}_j B^{sj} (1-B)^d (1-B^s)^D X_t^* = \varepsilon_t^*$, donde las primeras $d + s(D+P) + p$ observaciones $X_1^*, X_2^*, \ldots X_{d+s(D+P)+p}^*$ son extraídas con igual probabilidad de todos los $T - (d + s(D+P) + p + 1)$ posibles bloques de observaciones consecutivas de la serie original.

En la practica se genera una remuestra $SARIMA(p,d,0)(P,D,0)_S$ usando la recursion en el paso 7, con un tamaño de muestra igual a $T+100+10s$ y luego se descarta las primeras $100+10s$ observaciones en orden, para minimizar el efecto de valores iniciales.

En el siguiente paso se estima los parámetros de la serie generada por el *bootstrap* en el paso 6. Esto permite introducir la variabilidad de la estimación de los parámetros, sobre la estimación de los intervalos de predicción.

8. Dado $\{X_1^*, X_2^*, ..., X_T^*\}$ de los pasos previos, obtenemos la serie diferenciada *bootstrap* $\{Y_{d+sD+1}^*, Y_{d+sD+2}^*, ..., Y_T^*\}$ y se estiman los coeficientes autorregresivos $\hat{\boldsymbol{\phi}}_p^* = (\hat{\phi}_1^*, \hat{\phi}_2^*, ..., \hat{\phi}_p^*)'$ y $\hat{\boldsymbol{\Phi}}_p^* = (\hat{\Phi}_1^*, \hat{\Phi}_2^*, ..., \hat{\Phi}_P^*)'$ como en el paso 3.

En los pasos del 1. al 8. no es efectivo para la construcción de intervalos de predicción, por que el algoritmo no replica la distribución condicional de X_{T+h} dado la serie observada. Pero si fijamos las últimas $p+d+s(P+D)$ observaciones de la serie, se pueden obtener remuestras de los valores futuros X_{T+h}^*, dado $X_{T-(p+d+s(D+P))+1}^* = X_{T-(p+d+s(D+P))+1}, ..., X_T^* = X_T$.

9. Calculamos las observaciones *bootstrap* futuras mediante la recursión: $\sum_{j=0}^{p} \hat{\phi}_j^* B^j \sum_{j=0}^{P} \hat{\Phi}_j^* B^{sj} (1-B)^d (1-B^s)^D X_t^* = \varepsilon_t^*$, para $t = T+1, T+1, ..., T+h$ donde $h > 0$, y $X_t^* = X_t$, para $t \le T$.

10. Finalmente, la distribución *bootstrap* $F_{X_{T+h}^*}^*$, de X_{T+h}^*, es usado para aproximar la distribución desconocida de X_{T+h}, dado la serie observada. El estimado *Monte Carlo*

$\hat{F}^{*}_{X^{*}_{T+h}}$, es obtenido mediante una repetición en los pasos del (6) al (9), *B* veces. El intervalo de predicción estimado del $(1-\alpha)\%$ para X_{T+h}, esta dado por $\left[Q^{*}(\alpha/2), Q^{*}(1-\alpha/2)\right]$, donde $Q^{*}(\cdot) = \hat{F}^{*-1}_{X^{*}_{T+h}}$ son la distribución de las cuantilas estimadas mediante el *sieve bootstrap*.

5 RESULTADOS Y DISCUSIONES

En este capitulo se evalúa el desempeño del *Sieve Bootstrap*, para la construcción de intervalos de confianza de las predicciones a diferentes desplazamientos hacia el futuro, con diferentes números de replicas y diferentes criterios de selección de modelos autorregresivos de las series de tiempo de nubosidad, finalmente se comparará el *Sieve Bootstrap* con las técnicas convencionales.

5.1 Conjuntos de Datos

Las bases de datos para el presente análisis provienen de sensores ubicados en satélites, que han observado el comportamiento de las nubes a nivel global durante los últimos 21 años. Así se ve el desempeño del *Sieve Bootstrap* sobre 20 conjuntos de datos, listados en la tabla 5.2. Todos estos conjuntos de datos pertenecen al *International Satellite Clouds Climatology Product* (ISCCP).

5.1.1 Historial de los Datos

El *World Climate Research Program*, se plantea un problema en el cual está inmerso el estudio descriptivo y analítico de las nubes, diferentes tipos y aspectos referentes a ellas, como también la incidencia que tienen en otros fenómenos climatológicos, como la

precipitación de lluvia, temperatura de ambiente, etc. Así diferentes instituciones y organizaciones internacionales como NOAA-USA, EUMETSAT-EUROPA, AES-CANADA, JMA-JAPON, CMS-FRANCIA, INSAT-INDIA en su afán de realizar nuevas investigaciones, ponen a disposición a la comunidad científica un primer proyecto llamado *The International Satellite Cloud Climatology Project* (ISCCP), en el cual participan satélites como NOAA/TIROS-N, GOES-EAST, GOES-WEST, METEOSAT, GMS, INSTAT; dicho proyecto establecido en 1982 [Schiffer and Rossow, 1983], han establecido los siguientes objetivos:

1. Producir un conjunto de datos, de cobertura global de reflectancia, calibrada y normalizada, que contiene información de las propiedades de la atmósfera, del cual patrones de las nubes pueden ser derivadas.
2. Estimular a los investigadores para producir técnicas, algoritmos para inferir y validar propiedades físicas de las nubes
3. Proponer a los investigadores usar los datos de ISCCP, que contribuyan para la implementación y entendimiento de la radiación de la tierra y su ciclo hidrológico.

Este proyecto recolectó y analizó las medidas de la reflectancia de cinco satélites geoestacionarios y dos de orbita polar, para inferir la distribución global de las propiedades diarias, estacionales y variaciones interanuales de las nubes. El proyecto ISCCP proporciona información de nueve tipos de nubes figura (5.3) y las variables asociadas a ellas tales como temperatura, humedad, hielo/nieve de la superficie, etc. Estas últimas proveídas del proyecto TOVS, perteneciente a NOAA/NESDIS/NCDC. El método analítico para determinar la presencia o ausencia de nubes en un determinado elemento de imagen (*píxel*) se realiza

mediante propiedades radiométricas de las nubes para píxel nublado y píxel limpio; así el análisis del píxel funciona separadamente para cada conjunto de datos, es así que surgen diferentes niveles de datos DX, D1, D2, con diferentes tipos de resoluciones y diferentes tiempos de recolección.

Las fases de este proyecto son dos, C y D. Inicialmente en las fases C1 y C2 cubrían el periodo 07/1983-06/1991, cada tres horas y mensualmente; luego en las fase DX que comprende D1 y D2, cubren el periodo 07/1983 – 12/2004, cada tres horas y mensualmente, respectivamente.

Para el presente análisis se emplea la data D2, por la disponibilidad y sus diferentes características que ésta posee.

Los ejercicios desarrollados en este trabajo se apoyan en datos D2, principalmente por su disponibilidad.

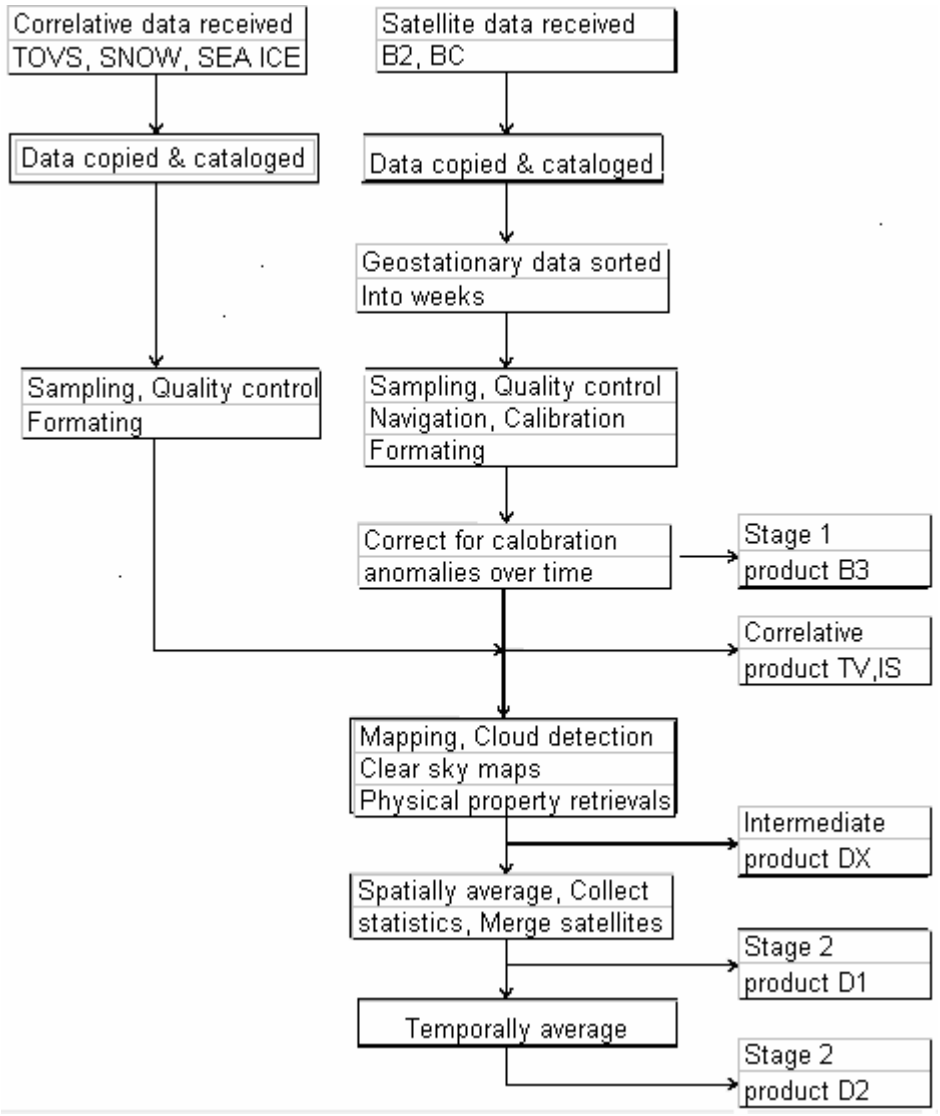


Figura 5.1. Esquema del procesamiento de la data D2 ISCCP.

5.1.2 Datos de Nivel D2

El archivo de datos de nivel D2 del ISCCP, tiene las siguientes características:

1. **Resolución:** $2.5^{\circ} \text{x} 2.5^{\circ}$, equivalente a 280x280 Km^2. Grid de área igual.
2. **Tiempo:** data mensual a nivel global.

3. **Contenido:** promedios mensuales de cantidades, incluidos los promedios diurnos, distribución y propiedades de nubosidad y tipos de nubes.

Los datos D2 del ISCCP, son observaciones mensuales en diferentes tiempos universales coordinados (UTC). El tiempo universal coordinado, es la zona horaria de referencia, respecto a la cual se calculan todas las otras zonas del mundo. El UTC es muy similar al *Greenwhich Mean Time* (GMT), la diferencia es que el UTC se sincroniza con el día y la noche del tiempo universal, al que se le añaden o quitan segundos de salto, tanto a finales de junio como en diciembre, cuando sea necesario. De esta forma los datos ISCCP D2 son recolectados en los siguientes tiempos universales coordinados:

- UTC=0, UTC=3, UTC=6, UTC=9, UTC=12, UTC=15, UTC=18, UTC=21 y el promedio de todos los anteriores UTC=*all*.

Los datos proveen información para cada grid. El globo terráqueo ha sido dividido en 6596 recuadros grids, cada uno con igual área, los recuadros están secuencialmente etiquetados comenzando en el polo sur y el meridiano de *Greenwhich*, procediendo hacia el este cubriendo 360^{o} de longitud, luego hacia el norte a la siguiente zona de latitud, existen 72 bandas de latitud y hasta un máximo de 144 bandas de longitud, formando así los recuadros grids de $2.5^{\circ} \times 2.5^{\circ}$, como se muestra en la figura 5.2.

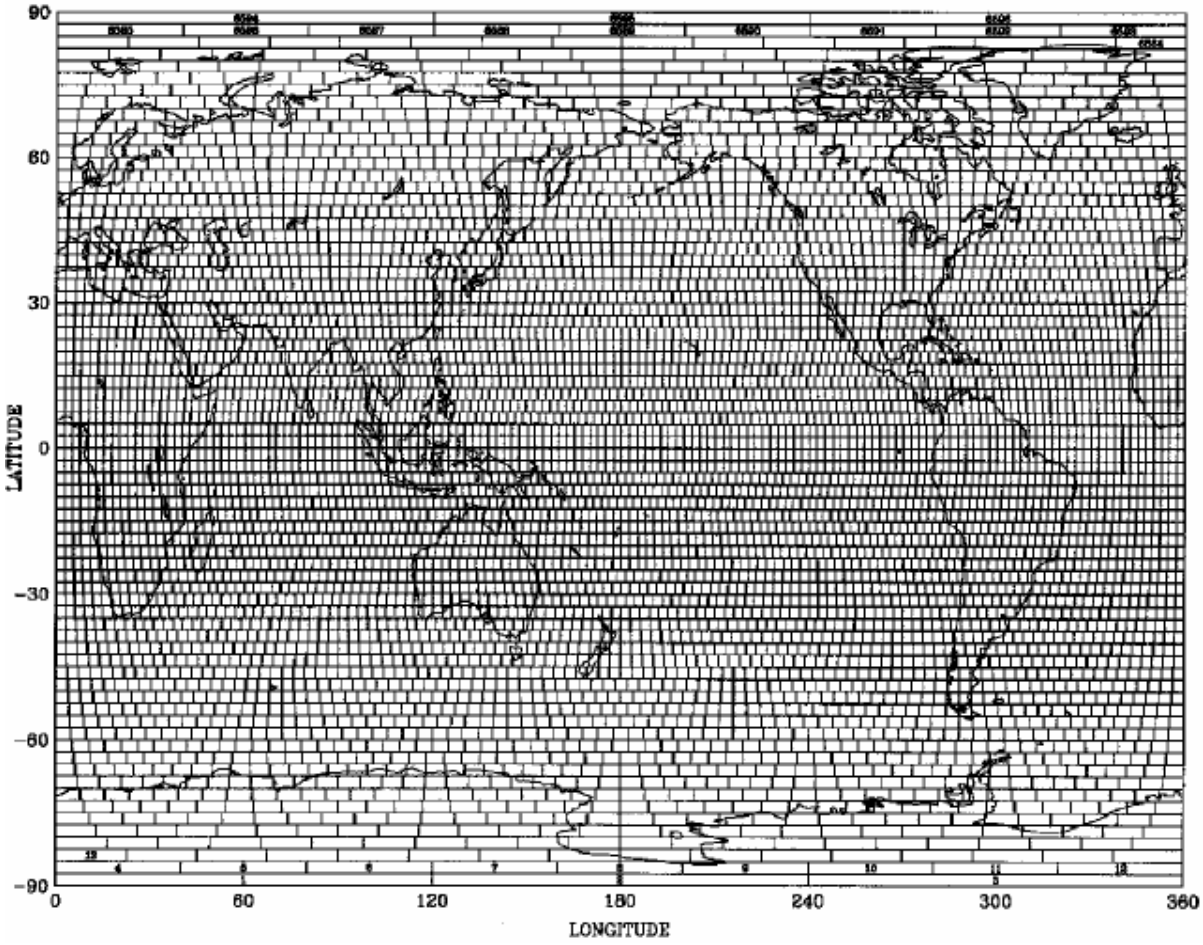


Figura 5.2. Mapa de recuadros grids de área igual para D2 data ISCCP a nivel globo.

El conjunto de datos está arreglada cronológicamente, con nueve archivos para cada mes, esto es, la data está separada cada tres horas en cada archivo 0, 3, 6,…,21, y el último archivo representa el promedio de los anteriores. Finalmente, la base de datos contiene en total 130 variables, distribuidas entre la identificación del recuadro, variables referentes a las nubes, tipos de nubes clasificadas por la presión atmosférica y densidad, clases de nubes clasificadas por sus aspectos y estados, así como variables relacionadas a la atmósfera.

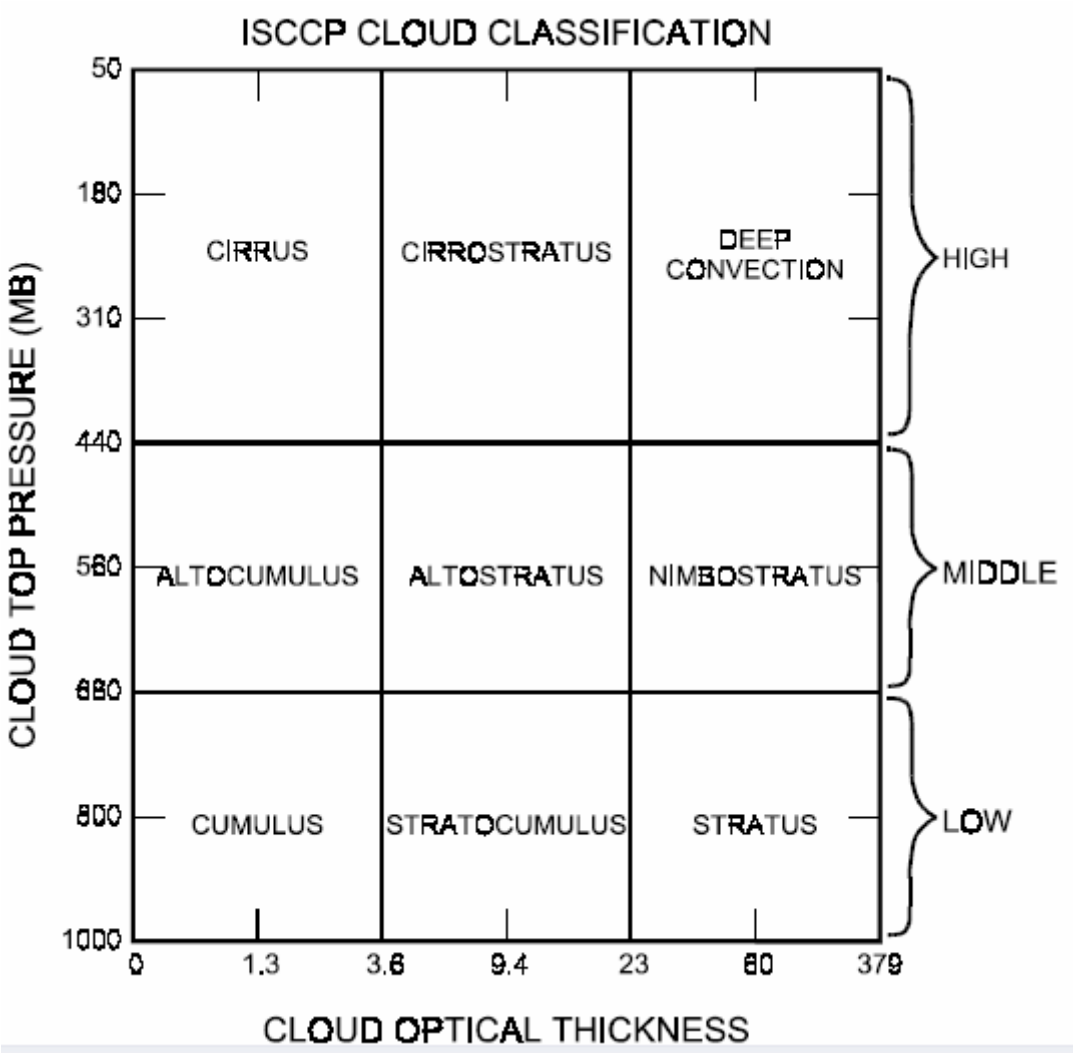


Figura 5.3. Clasificación de las Nubes según ISCCP.

Finalmente para la adquisición de la data D2, se procede con la lectura de los archivos comprimidos existentes en los productos CD's, mediante el programa FORTRAN, y subrutinas para la lectura, decodificación y uso de la data D2; para ello, se proporciona un programa SAMPLE que se presenta en el apéndice.

Mediante todas estas subrutinas y el programa SAMPLE, podemos extraer la data D2, delimitada para cualquier área del globo.

5.1.3 Segmentación de los Datos de Nivel D2 Según el Área de Trabajo

La segmentación del área de trabajo para el presente análisis es como sigue:

La zona del Caribe, mediante las cuatro islas representativas: Cuba, La Española (República Dominicana y Haití), Jamaica y Puerto Rico, ubicadas dentro del globo: Latitud 15N – 25N, Longitud 63W – 87W aproximadamente y el periodo de referencia para el análisis es desde Julio/1983, hasta Diciembre/2004, finalmente el número de recuadros grids son aproximadamente 36, esto es que tenemos 36 estaciones ubicadas en el área de trabajo.

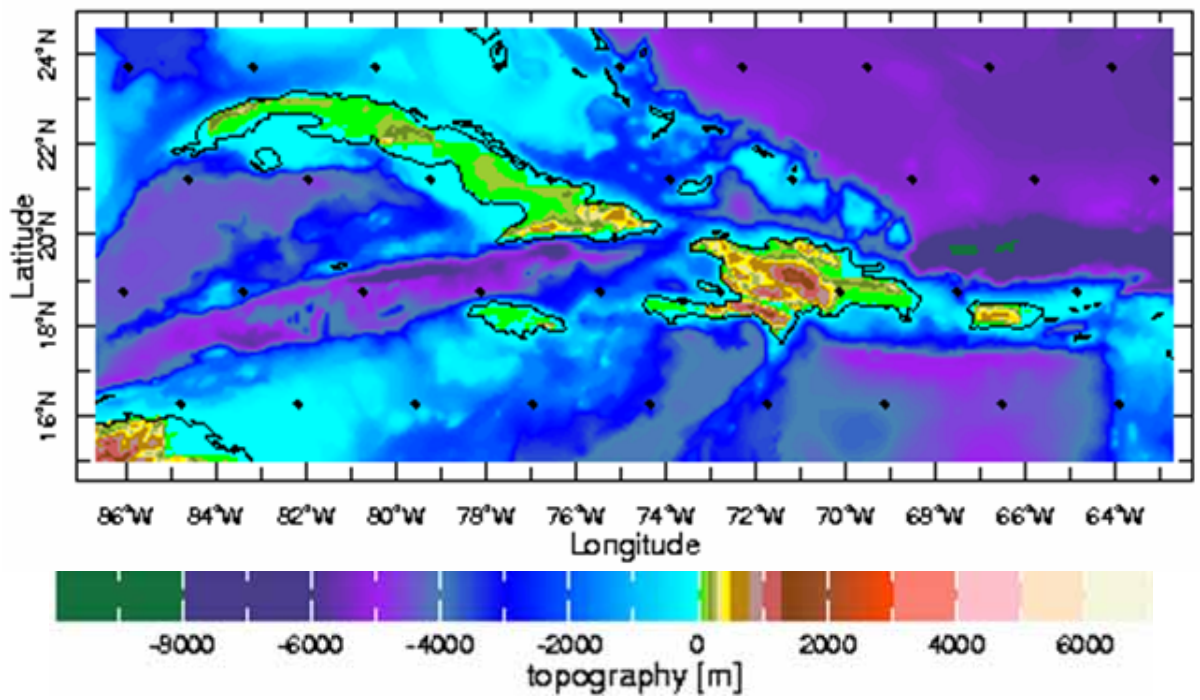


Figura 5.4. Area de Trabajo, Islas Representativas del Caribe y Estaciones del ISCCP.

Para el presente análisis consideramos el promedio del tiempo diurno en las cuatro islas representativas del Caribe: Cuba, La Española (República Dominicana y Haití), Jamaica y Puerto Rico. Este promedio es tomado considerando el tiempo diurno en Puerto Rico y Cuba, ya que son los extremos en la longitud del área de trabajo, esto es:

Tabla 5.1 Tiempo Universal Coordinado en las Longitudes Extremas del área de trabajo

Tiempo Diurno		
Tiempo Universal Coordinado (UTC)	*Tiempo Diurno en Puerto Rico*	*Tiempo Diurno en Cuba*
UTC = 12	8:00 AM	7:00 AM
UTC = 15	11:00 AM	10:00 AM
UTC = 18	2:00 PM	1:00 PM
UTC = 21	5:00 PM	4:00 PM

- **Area de Trabajo**:

Islas representativas del Caribe: Cuba, La Española (República Dominicana & Haití), Jamaica y Puerto Rico.

- **Ubicación**

Latitud 15N – 25N, Longitud 63W – 87W

- **Intervalo de Tiempo**

Julio/1983 - Diciembre/2004.

5.1.4 Variables de los Datos de Nivel D2

Para probar la metodología del *Sieve Bootstrap* y construir intervalos de predicción para series de tiempo, se considera las **Series de Tiempo de Nubosidad**. Estas series de tiempo

referentes a las nubes, son en realidad diferentes realizaciones de procesos para diferentes tipos de nubes que presentan una estrecha relación con la precipitación de lluvia, por esta razón se detalla a continuación las cualidades de los datos de nubes que analizaremos.

Tipos de Nubes por su Forma Consideradas en el Analisis:

1. Estratocúmulos (Stratocumulus).
2. Estratos (Stratus).
3. Nimbo Estratos (Nimbostratus).
4. Nubes de Desarrollo Vertical (Deep Convection).

Variables en los Diferentes Tipos de Nubes según su Forma Consideradas en el Analisis:

- Cantidad de las nubes (Amount or Cloud cover).

Unidad de medida: Porcentaje (%).

- Presión en la parte superior de las nubes (Top Pressure).

Unidad de medida: Milibares (Mb).

- Temperatura en la parte superior de las nubes (Top Temperatura).

Unidad de medida: Grados Kelvin (K).

- Densidad óptica de las nubes (Optical Thickness).

Unidad de medida: Tau (TAU).

- Agua precipitable de las nubes (Water Path).

Unidad de medida: Gramos sobre metros cuadrados (g/m^2).

De esta forma se generan 20 conjuntos de series de tiempo para el análisis siguiente.

Tabla 5.2 Variables para el Análisis del Sieve Bootstrap en Series de Tiempo Climatologicas.

Variables para los Cuatro Tipos de Nubes				
Nro	*Tipos de Nubes por su forma*	*Variables*	*Abreviaciones usadas*	*Unidades*
1	Stratocumulus	Amount	SCA	(%)
2	Stratocumulus	Top Pressure	SCTP	(Mb)
3	Stratocumulus	Top Temperature	SCTT	(K)
4	Stratocumulus	Optical Thickness	SCOT	(TAU)
5	Stratocumulus	Water Path	SCWP	(g/m)
6	Stratus	Amount	SA	(%)
7	Stratus	Top Pressure	STP	(Mb)
8	Stratus	Top Temperature	STT	(K)
9	Stratus	Optical Thickness	SOT	(TAU)
10	Stratus	Water Path	SWP	(g/m)
11	Nimbostratus	Amount	NSA	(%)
12	Nimbostratus	Top Pressure	NSTP	(Mb)
13	Nimbostratus	Top Temperature	NSTT	(K)
14	Nimbostratus	Optical Thickness	NSOT	(TAU)
15	Nimbostratus	Water Path	NSWP	(g/m)
16	Deep Convection	Amount	DCA	(%)
17	Deep Convection	Top Pressure	DCTP	(Mb)
18	Deep Convection	Top Temperature	DCTT	(K)
19	Deep Convection	Optical Thickness	DCOT	(TAU)
20	Deep Convection	Water Path	DCWP	(g/m)

5.2 Procesamiento

En las metodologías discutidas en el presente trabajo, se asume que todas las series de tiempo de nubosidad son continuas, medidas en tiempo discreto; y por ser unidimensionales, representan el centroide (promedio) del área de trabajo. En la obtención de los datos se presentaron valores perdidos en los tiempos diurnos dentro de las diferentes estaciones en el área de trabajo; para remediar este problema se usa un algoritmo de interpolación *Kriging* ,

en tiempo y espacio utilizando los *6* vecinos más cercanos, de esta manera se soluciona el problema de los valores perdidos que no generan problema alguno debido a que se trabaja con el promedio. Por otro lado Alonso (2004), genera una rutina en lenguaje FORTRAN para la construcción de intervalos de predicción, así mismo rutinas en MATLAB, para la selección de modelos autorregresivos. En el presente trabajo se elaboró rutinas en MATLAB, para el editado de la data, construcción de mapas, análisis descriptivo y análisis mediante el *Sieve Bootstrap* y *Box Jenkins*.

Modelamiento Box y Jenkins:

Previamente consideramos procesos estacionarios autorregresivos y medias móviles con componente estacional, para aproximar nuestras series de tiempo de nubosidad mediante la metodología *Box y Jenkins*, así se describe paso a paso este tratado.

- **Visualización:** Se presentan las 20 series de tiempo de nubosidad descritas en la tabla 5.2. La figura 5.5 presenta a las series de tiempo de nubosidad, las cuales presentan en su estructura el componente estacional habitual en las series de tiempo climatologicas, excepto en la serie SCTP; el componente de tendencia en algunas de las series, y el componente estocástico presente en todas las series de nubosidad.

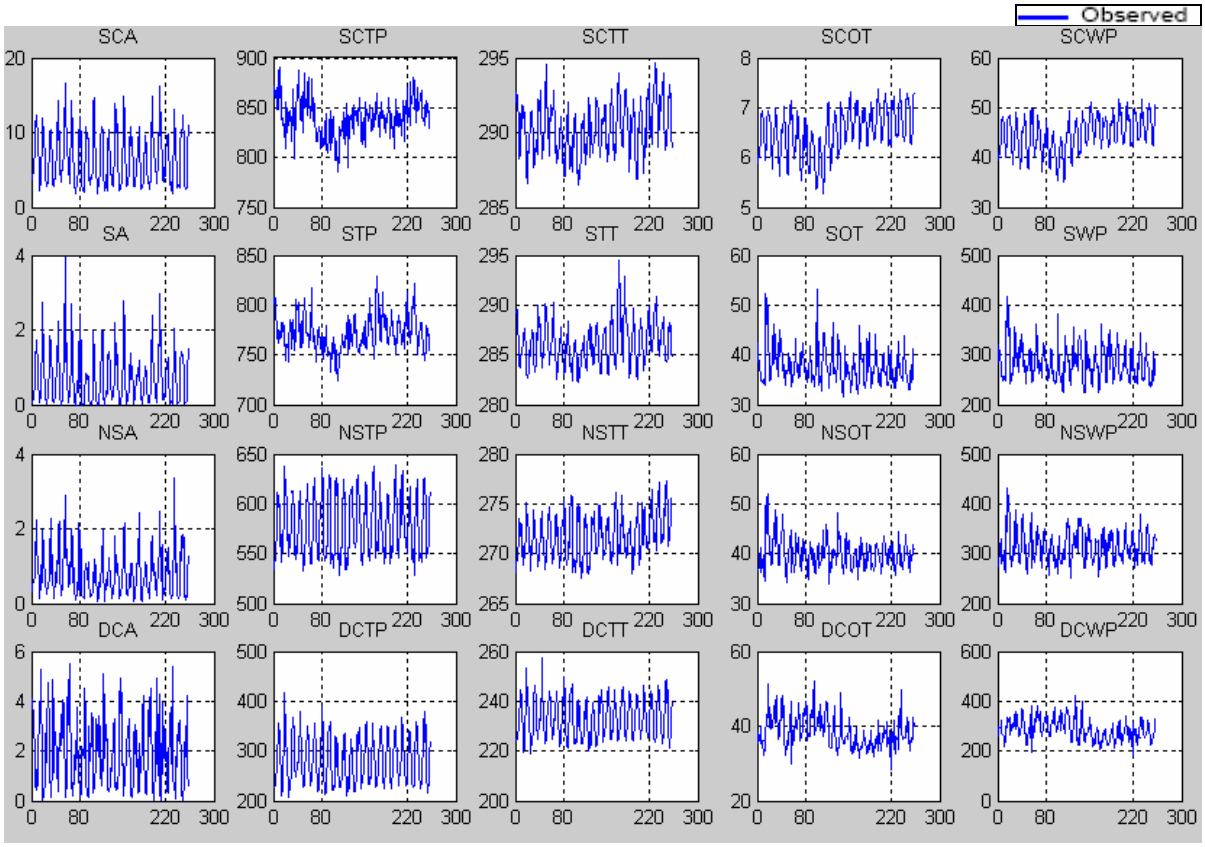


Figura 5.5: Series de Tiempo de Nubosidad, periodo julio/83-diciembre/04.

- **Identificación:** Las 20 series de tiempo son identificadas utilizando las diferentes técnicas descritas en el ciclo interactivo de *Box Jenkins*. La función de autocorrelación muestral presentada en la figura 5.6 reafirma la presencia del componente estacional *s* con amplitud igual a 12, excepto en la serie SCTP. De esta manera las 19 series de tiempo de nubosidad presentan una estacionalidad $s = 12$, y la serie excluida de estacionalidad SCTP, no es estacionario.

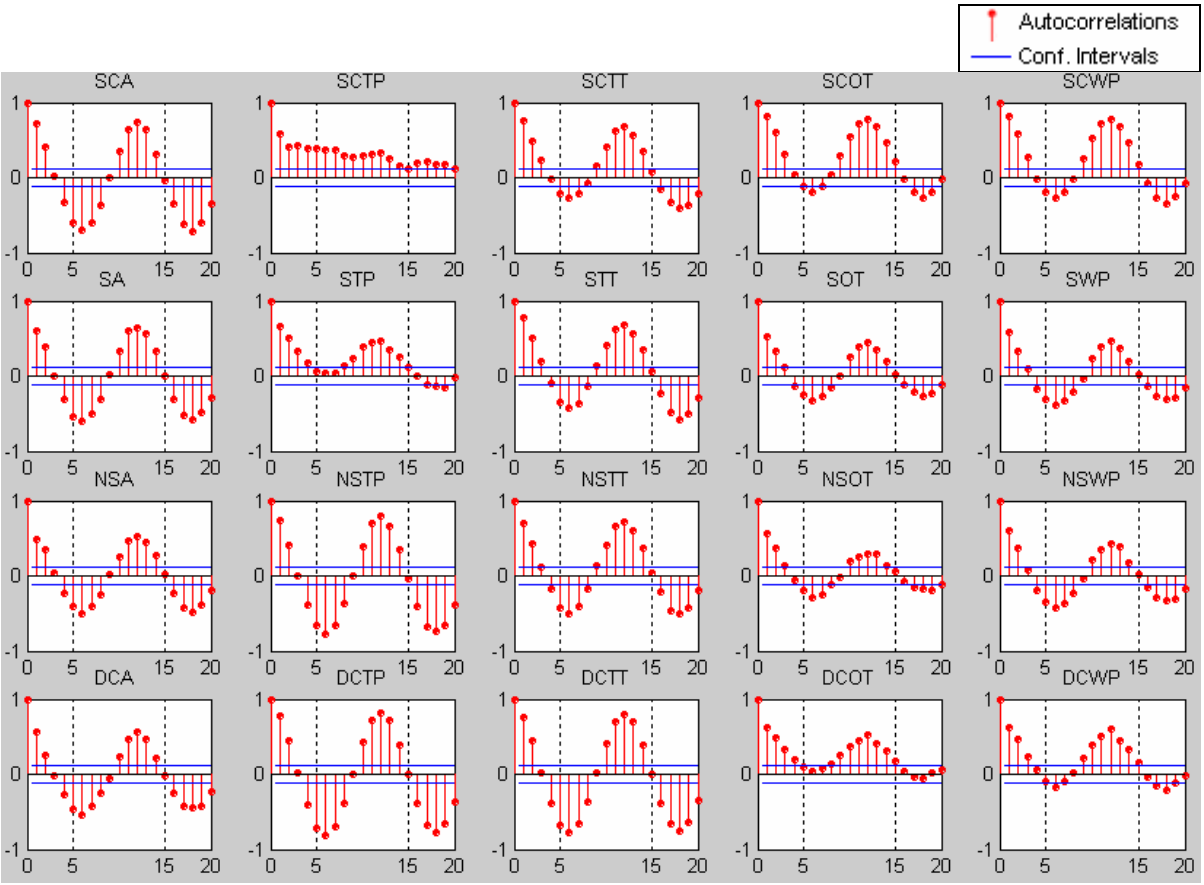


Figura 5.6: Funciones de Autocorrelación Muestral de todas las Series de Tiempo de Nubosidad.

Utilizando procedimientos de identificación convencionales y alternativos para conocer los órdenes de los modelos $SARIMA(p,d,q)(P,D,Q)_S$, de la forma: $\phi(B)\Phi(B^s)\Delta^d\Delta_s^D(X_t-\mu)=\theta(B)\Theta(B^s)\varepsilon_t$ descritos en (2.24), para las 20 series de tiempo de nubosidad, se obtienen los resultados generales, descritos en la tabla 5.3, donde se presenta lo siguiente:

Las series de tiempo necesitan una transformación adecuada *Box y Cox*, descrita en (2.26), mediante el parámetro de transformación λ.

Las series de tiempo posiblemente necesitan una constante (media) en el modelo.

Las series de tiempo necesitan ser estacionarias, para ello se toman diferencias regulares o estacionales, o ambas según sea caso.

Tabla 5.3: Identificación de Modelos SARIMA(p,d,q)(P,D,Q)$_S$ de todas las Series de Tiempo de Nubosidad.

MODELOS BOX JENKINS									
Series	*Transformación Box & Cox*	*Mean*	*p*	*d*	*q*	*P*	*D*	*Q*	*RMSE*
SCA	0	0	0	1	1	0	1	1	1.58153
SCTP	1	0	0	1	2	0	0	0	14.6511
SCTT	1	0	0	1	1	0	1	1	0.88069
SCOT	1	0	0	1	1	0	1	1	0.16987
SCWP	1	0	0	1	1	0	1	1	1.33663
SA	1	0	0	0	2	0	1	1	0.37945
STP	0	0	0	1	1	0	1	1	11.768
STT	0	0	0	1	1	0	1	1	1.04657
SOT	0	0	0	1	2	0	1	1	2.69102
SWP	0	1	0	0	2	0	1	1	23.3152
NSA	1	0	0	1	1	0	1	1	0.40957
NSTP	0	0	0	1	1	0	1	1	11.794
NSTT	1	1	0	1	1	0	1	1	1.05618
NSOT	0	1	0	0	2	0	1	1	2.2271
NSWP	0	0	0	0	2	0	1	1	21.2672
DCA	1	1	0	0	1	0	1	1	0.8253
DCTP	0	0	0	1	1	0	1	1	17.3297
DCTT	0	0	0	1	1	0	1	1	3.20111
DCOT	0	0	0	1	1	0	1	1	2.81211
DCWP	0	0	0	1	1	0	1	1	27.6091

Finalmente los ordenes de las series de tiempo estacionarias p, q, P, Q, se identifican mediante el estudio de la función de autocorrelación y autocorrelación parcial, muestrales de

las series estacionarias así como también un criterio de selección alternativo como es la raíz del error cuadrático promedio *RMSE.*

- **Estimación:** Luego de la identificación de los órdenes, y utilizando un apropiado método apropiado de estimación (Mínimos Cuadrados Condicionales), se estiman los parámetros de los modelos identificados en la tabla 5.3.

Tabla 5.4: Estimación de Parámetros SARIMA(p,d,q)(P,D,Q)$_S$ de todas las Series de Tiempo de Nubosidad.

COEFICIENTES ARMA REGULARES y ESTACIONALES				
Series	*theta_1*	*theta_2*	*THETA_1*	*Constante*
SCA	0.65942	0	0.933631	0
SCTP	0.483747	0.301647	0	0
SCTT	0.674606	0	0.943611	0
SCOT	0.565603	0	0.942255	0
SCWP	0.579709	0	0.942879	0
SA	-0.108082	-0.258708	0.9653	0
STP	0.663097	0	0.934317	0
STT	0.619324	0	0.946321	0
SOT	0.727395	0.170297	0.93327	0
SWP	-0.302945	-0.154565	0.93214	-0.0035387
NSA	0.90627	0	0.910036	0
NSTP	0.765554	0	0.913421	0
NSTT	0.857827	0	0.909535	0.00223399
NSOT	-0.355092	-0.239108	0.943393	-0.0016148
NSWP	-0.331333	-0.238813	0.949522	0
DCA	-0.194124	0	0.938111	-0.024823
DCTP	0.813962	0	0.92056	0
DCTT	0.843091	0	0.920308	0
DCOT	0.703155	0	0.934287	0
DCWP	0.761882	0	0.937228	0

- **Diagnóstico:** Esta fase esta dedicada íntegramente a los residuales de los modelos ajustados en las fases anteriores, los resultados se presentan en la tabla 5.5, donde se presentan los promedios de los residuales, los promedios de los errores absolutos *MAE*, los errores estándar estimados de los residuales *SE(res)*, los grados de libertad efectivos para estimar los errores estándar, los estadísticos *Box-Pierce* para las primeras 24 autocorrelaciones *Q(24)*, y sus respectivos *P-valores*.

Tabla 5.5: Diagnóstico de Modelos SARIMA(p,d,q)(P,D,Q)$_S$ de todas las Series de Tiempo de Nubosidad.

ESTADISTICOS DE LOS RESIDUOS						
Series	*Mean*	*MAE*	*SE(res)*	*DF*	*Box-Pierce Q(24)*	*P_valor*
SCA	0.05475	1.14231	0.23601	243	28.7114	0.15329
SCTP	-0.35003	11.5854	14.6513	255	25.8374	0.25873
SCTT	0.00875	0.7009	0.89912	243	25.8051	0.26015
SCOT	0.00014	0.13516	0.17399	243	33.2874	0.05794
SCWP	0.00027	1.06318	1.3701	243	31.0497	0.0951
SA	-0.00753	0.24714	0.3832	243	21.0198	0.45773
STP	0.4619	9.00028	0.01528	243	18.5771	0.67127
STT	0.02589	0.73857	0.00368	243	25.4131	0.27773
SOT	0.17031	1.84516	0.06723	242	20.2903	0.50295
SWP	2.66049	17.3313	0.07955	242	23.9006	0.2979
NSA	0.03596	0.28577	0.41491	243	12.7756	0.93923
NSTP	-0.50474	9.08256	0.02044	243	17.929	0.71019
NSTT	-0.05831	0.80736	1.06857	242	15.8947	0.82108
NSOT	0.13168	1.61546	0.05591	242	21.6767	0.41834
NSWP	1.51799	16.0658	0.06814	243	22.5822	0.36665
DCA	0.02792	0.6194	0.83843	243	27.355	0.19805
DCTP	-0.3773	12.9435	0.05938	243	22.7359	0.41677
DCTT	-0.07645	2.43315	0.01385	243	21.6086	0.48345
DCOT	0.02222	2.07039	0.07261	243	20.7261	0.53772
DCWP	0.18699	20.3486	0.09706	243	27.2	0.20373

Por otro lado en la Figura 5.7 se muestra los histogramas de los residuales, para los modelos ajustados, presentando muchos de estos diferentes tipos de asimetrías y apuntamiento.

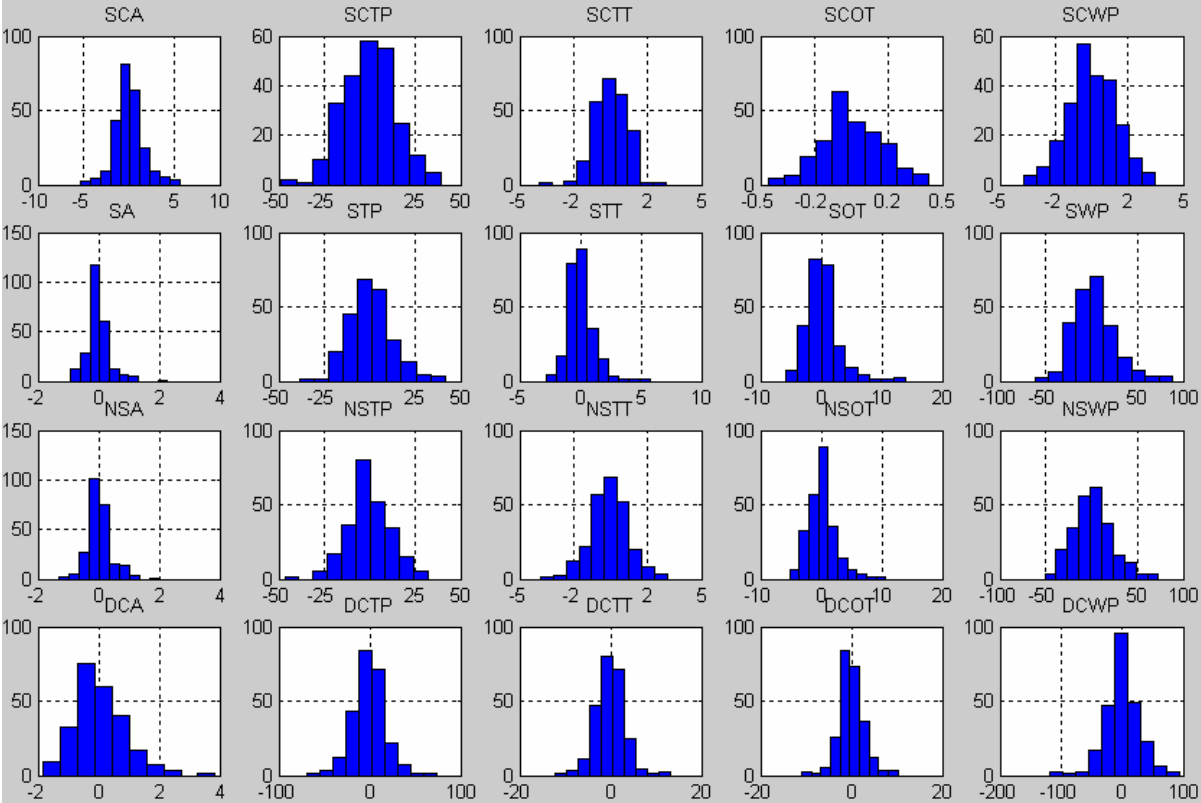


Figura 5.7: Histogramas de Residuos de los Modelos Ajustados para las Series de Tiempo de Nubosidad.

Para comprobar si verdaderamente los residuales no están distribuidos según una distribución normal, en la tabla 5.6 nos muestra la prueba de normalidad de *Jarque Bera*, donde la hipótesis nula se refiere a que los datos siguen una distribución normal, mientras que la alternativa dice lo contrario.

Tabla 5.6: Prueba de Normalidad de Jarque Bera para los Modelos SARIMA(p,d,q)(P,D,Q)$_S$

Series	Jarque Bera	P-valor
SCA	34.7595	2.83E-08
SCTP	0.6764	0.7131
SCTT	14.529	7.00E-04
SCOT	0.1268	0.9386
SCWP	0.2091	0.9007
SA	453.7886	0
STP	10.6849	0.0048
STT	304.186	0
SOT	374.5997	0
SWP	41.2495	1.10E-09
NSA	123.6155	0
NSTP	4.5323	0.1037
NSTT	4.1566	0.1251
NSOT	108.6005	0
NSWP	14.3977	7.47E-04
DCA	53.3117	2.65E-12
DCTP	41.1763	1.14E-09
DCTT	35.4025	2.05E-08
DCOT	47.4756	4.91E-11
DCWP	41.6366	9.09E-10

Los residuos de las series SCTP, SCOT, SCWP, NSTP y NSTT, siguen una distribución normal bajo la prueba de *Jarque Bera* con un nivel de significación del 5%.

La metodología *Box Jenkins* mostrada para las 20 series de tiempo representan modelos adecuados para describir los datos, consecuentemente bajo el ciclo iterativo se pueden obtener predicciones *h* pasos adelante, donde $h \in \mathbb{N}$. Usualmente como se describe en la metodología *Box Jenkins*, estos modelos son buenos predictores a periodos cortos, por tal motivo se debe escoger un *h* adecuado para las predicciones. No obstante previamente se realiza una validación para los modelos propuestos en la tabla 5.3, tomando como muestra de

entrenamiento $n = 246$, que comprende el periodo Julio/1983 hasta Diciembre/2003; y como muestra de prueba $n = 12$ que comprende el periodo Enero/2004 hasta Diciembre/2004.

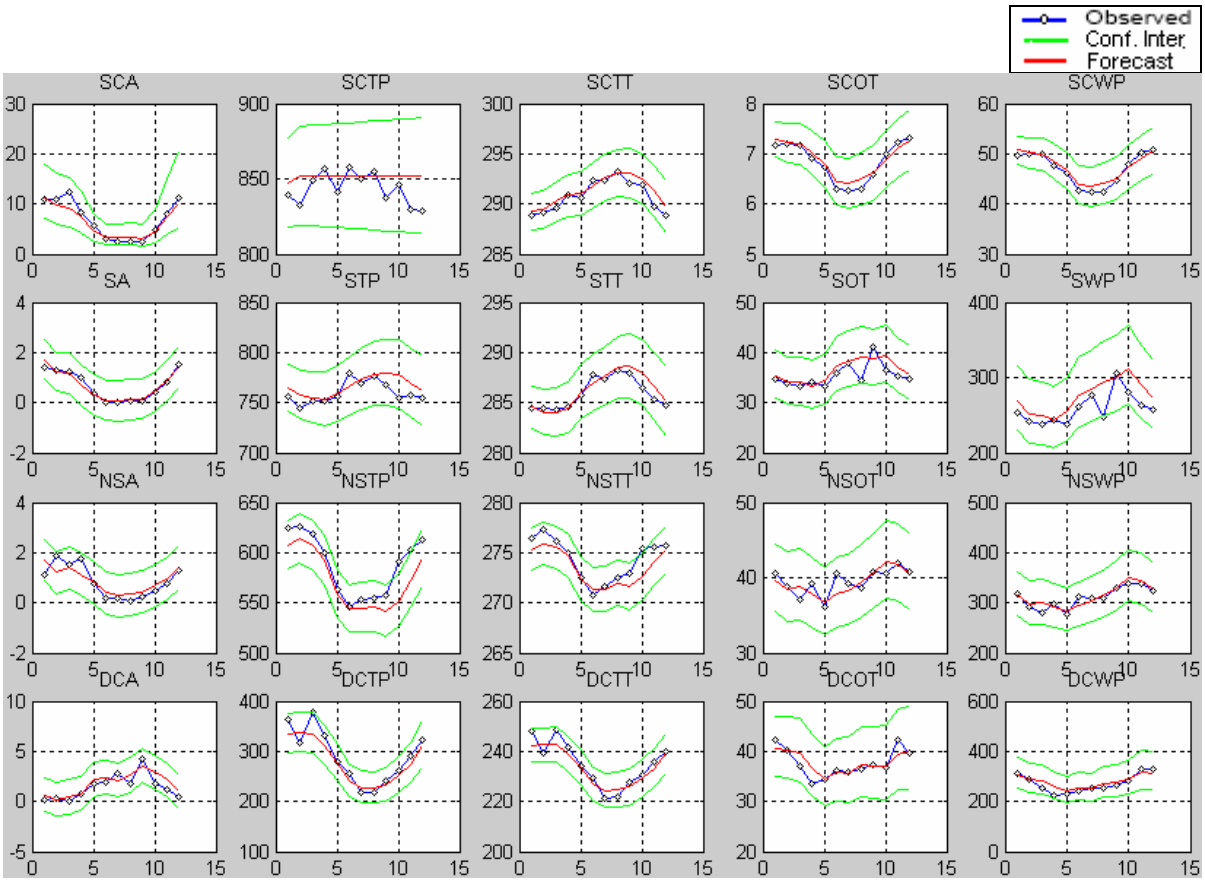


Figura 5.8: Validación para las ultimas 12 Observaciones de los Modelos Ajustados para las Series de Tiempo de Nubosidad.

En la figura 5.8 se muestra la validación para las últimas doce observaciones, la figura comprende los datos observados para el periodo Enero/2004 a Diciembre/2004, sus respectivas predicciones con sus correspondientes intervalos de predicción al 5% de significación, bajo los modelos ajustados mediante la metodología *Box Jenkins*

Finalmente el objetivo principal de la metodología *Box Jenkins*, el de realizar predicciones, es desarrollada para doce periodos adelante.

- **Predicciones para el año 2005 bajo los modelos Box Jenkins.**

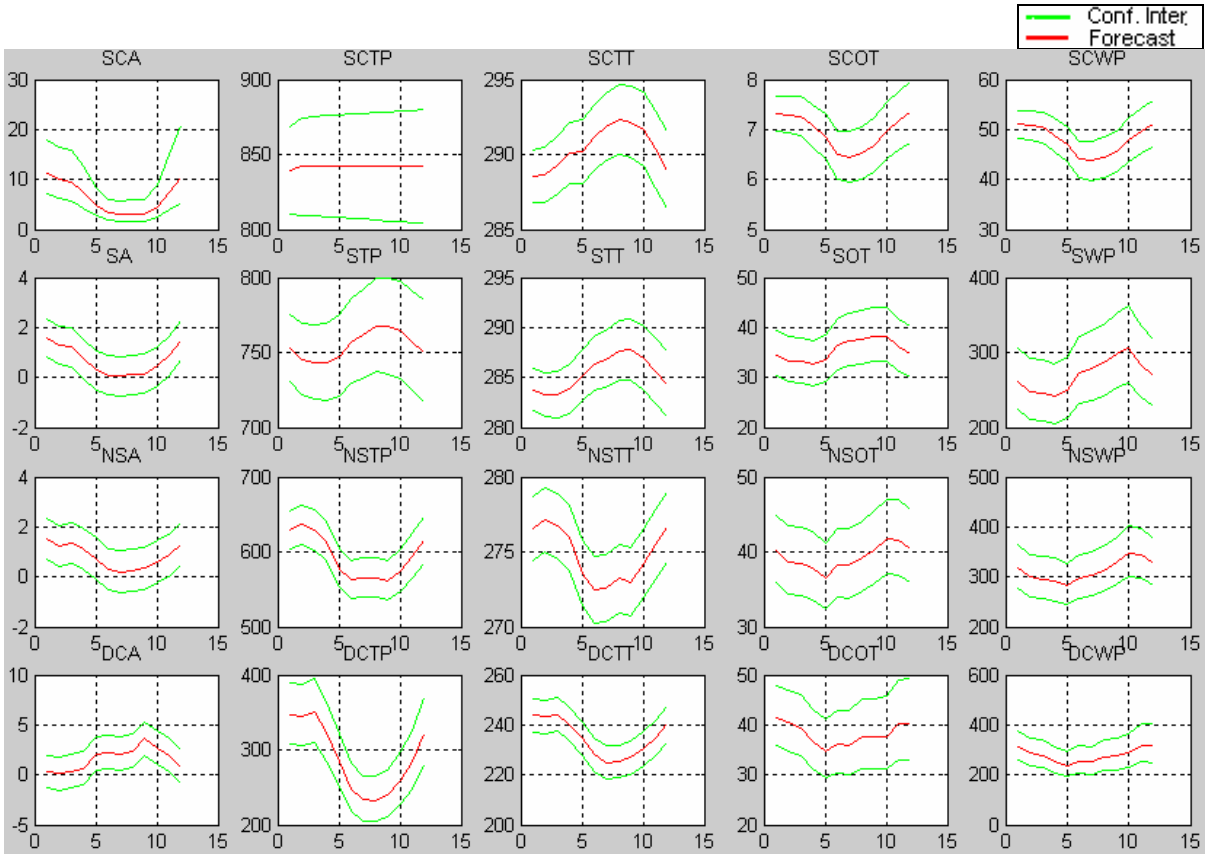


Figura 5.9: Predicciones de los Modelos Ajustados para las Series de Tiempo de Nubosidad, para 12 periodos adelante Enero/2005 a Diciembre/2005.

En la figura 5.9 se realizan las predicciones con sus respectivos intervalos de confianza con un nivel de significancia del 5%, para los periodos de Enero/2005 a Diciembre/2005, bajo los modelos ajustados, estas predicciones aparentemente siguen el patrón estacional.

Aproximación Sieve Bootstrap:

El método de *sieve bootstrap*, mostrado a continuación presenta el desarrollo práctico descrito en el capitulo 4, mediante la elaboración de aproximaciones de procesos autorregresivos para las series de tiempo de nubosidad, en sus fases de identificación de órdenes, selección de modelos, estimación de parámetros, construcción de distribuciones empíricas de los residuales, remuestreo en los residuales, construcción de nuevos estimadores y finalmente generación de predicciones e intervalos de predicción.

- **Desarrollo del Sieve Bootstrap en las Series de Tiempo de Nobusidad para Predicciones e Intervalos de Predicción:**

Asumiendo dos criterios de selección de modelos autorregresivos: Criterio de Información de *Akaike* (AIC), y Criterio Corregido de Información de *Akaike* (AICC). Por otro lado en la etapa de remuestreo se toman tres diferentes números de remuestras *bootstraps*: $B = 200$, $B = 1000$ y $B = 2000$ del mismo tamaño . Finalmente se toman doce predicciones adelante, en la etapa de predicción y doce en la etapa de validación, esto es $h = 1, 6, 12$.

- **Identificación de modelos con el AIC y AICC, bajo el Sieve Bootstrap.**

Tabla 5.7: Identificación de Modelos Sieve Bootstrap, con el AIC.

MODELOS SIEVE BOOTSTRAP ,AIC,B=200,1000,2000,h=1,6,12									
Series	*Transformación Box & Cox*	*Mean*	*p*	*d*	*q*	*P*	*D*	*Q*	*AIC*
SCA	0	0	15	1	0	3	1	0	-2.8275
SCTP	1	0	15	1	0	0	0	0	5.3914
SCTT	1	0	4	1	0	3	1	0	-0.0873
SCOT	1	0	24	1	0	3	1	0	-3.4084
SCWP	1	0	4	1	0	3	1	0	0.73885
SA	1	0	24	0	0	3	1	0	-1.7954
STP	0	0	5	1	0	3	1	0	-8.1976
STT	0	0	7	1	0	3	1	0	-11.038
SOT	0	0	9	1	0	3	1	0	-5.23
SWP	0	1	2	0	0	3	1	0	-4.9765
NSA	1	0	23	1	0	2	1	0	-1.5499
NSTP	0	0	8	1	0	3	1	0	-7.6965
NSTT	1	1	10	1	0	3	1	0	0.21517
NSOT	0	1	2	0	0	3	1	0	-5.7526
NSWP	0	0	2	0	0	3	1	0	-5.3215
DCA	1	1	24	0	0	2	1	0	-0.2229
DCTP	0	0	11	1	0	3	1	0	-5.5477
DCTT	0	0	12	1	0	3	1	0	-8.4524
DCOT	0	0	24	1	0	3	1	0	-5.1153
DCWP	0	0	8	1	0	3	1	0	-4.5166

En esta etapa de identificación de los modelos, se consideran los dos criterios de elección AIC y AICC, en las tablas 5.7 y 5.8 se muestran los resultados en los cuales el AICC por su naturaleza tiene a corregir el sesgo que el AIC produce, seleccionando así ordenes un poco más reducidos.

Tabla 5.8: Identificación de Modelos Sieve Bootstrap, con el AICC.

MODELOS SIEVE BOOTSTRAP ,AICC,B=200,1000,2000,h=1,6,12									
Series	*Transformación Box & Cox*	*Mean*	*p*	*d*	*q*	*P*	*D*	*Q*	*AICC*
SCA	0	0	15	1	0	3	1	0	-1.8074
SCTP	1	0	15	1	0	0	0	0	6.4079
SCTT	1	0	4	1	0	3	1	0	0.97077
SCOT	1	0	4	1	0	3	1	0	-2.3892
SCWP	1	0	4	1	0	3	1	0	1.7488
SA	1	0	2	0	0	3	1	0	-0.7696
STP	0	0	5	1	0	3	1	0	-7.1871
STT	0	0	7	1	0	3	1	0	-10.026
SOT	0	0	9	1	0	3	1	0	-4.2164
SWP	0	1	2	0	0	3	1	0	-3.9675
NSA	1	0	6	1	0	3	1	0	-0.5246
NSTP	0	0	8	1	0	3	1	0	-6.6838
NSTT	1	1	10	1	0	3	1	0	1.2296
NSOT	0	1	2	0	0	3	1	0	-4.7436
NSWP	0	0	2	0	0	3	1	0	-4.3124
DCA	1	1	24	0	0	2	1	0	0.81037
DCTP	0	0	11	1	0	3	1	0	-4.5323
DCTT	0	0	12	1	0	3	1	0	-7.4359
DCOT	0	0	24	1	0	3	1	0	-4.0801
DCWP	0	0	8	1	0	3	1	0	-3.5039

- **Estimación de Parámetros bajo el Sieve Bootstrap.**

Tabla 5.9: Estimación de Parámetros de los Modelos Sieve Bootstrap con el AIC

COEFICIENTES AUTORREGRESIVOS REGULARES & ESTACIONALES CON SIEVE BOOTSTRAP ,AIC,B=200,1000,2000,h=1,6,12																				
Coef.	*SCA*	*SCTP*	*SCTT*	*SCOT*	*SCWP*	*SA*	*STP*	*STT*	*SOT*	*SWP*	*NSA*	*NSTP*	*NSTT*	*NSOT*	*NSWP*	*DCA*	*DCTP*	*DCTT*	*DCOT*	*DCWP*
Cte.	0	0	0	0	0	0	0	0	0	-0.003	0	0	0	-0.001	0	-0.02	0	0	0	0
phi_1	-0.6	-0.5	-0.6	-0.4	-0.5	0.028	-0.6	-0.5	-0.7	0.307	-0.9	-0.7	-0.8	0.3187	0.3	0.179	-0.7	-0.7	-0.5	-0.682
phi_2	-0.4	-0.6	-0.6	-0.31	-0.33	0.249	-0.4	-0.4	-0.5	0.151	-0.7	-0.5	-0.7	0.2374	0.26	0.003	-0.5	-0.5	-0.3	-0.469
phi_3	-0.4	-0.3	-0.2	-0.2	-0.24	0.002	-0.3	-0.3	-0.4		-0.6	-0.4	-0.6			-0.01	-0.3	-0.4	-0.1	-0.408
phi_4	-0.2	-0.3	-0.1	-0.23	-0.19	-0.04	-0.2	-0.2	-0.4		-0.6	-0.2	-0.4			-0.03	-0.4	-0.5	-0.1	-0.307
phi_5	-0.1	-0.2		-0.06		-0.02	-0.2	-0.2	-0.3		-0.6	-0.2	-0.4			-0	-0.4	-0.4	-0.1	-0.308
phi_6	-0.1	-0.2		-0.05		0.03		-0.1	-0.2		-0.5	-0.2	-0.4			-0.11	-0.2	-0.3	-0.2	-0.298
phi_7	-0.2	-0.1		-0.04		0.048		-0.1	-0.2		-0.4	-0.1	-0.3			0.038	-0.2	-0.3	-0.1	-0.231
phi_8	-0.3	-0.2		-0.08		-0.03			-0.1		-0.4	-0.1	-0.3			0.067	-0.3	-0.3	-0	-0.159
phi_9	-0.2	-0.2		-0.08		0.018			-0.2		-0.4		-0.2			-0.05	-0.3	-0.3	0.06	
phi_10	-0.1	-0.1		0.03		0.082					-0.3		-0.2			-0	-0.3	-0.3	0.03	
phi_11	-0.1	-0.1		0.05		0.047					-0.3					-0.04	-0.2	-0.2	-0	
phi_12	-0.6	0.09		-0.83		0.412					0.2					0.369		-0.1	-0.9	
phi_13	-0.3	0.07		-0.15		0.037					0.2					-0.03			-0.3	
phi_14	-0.2	0.02		-0.09		0.063					0.21					0.02			-0.2	
phi_15	-0.2	-0.1		-0.14		0.012					0.25					0.034			-0.1	
phi_16				-0.14		-0.04					0.25					-0.09			-0	
phi_17				-0.11		-0.02					0.18					0.013			-0.1	
phi_18				-0.16		0.011					0.17					0.042			-0.2	
phi_19				-0.12		0.036					0.26					0.008			-0.1	
phi_20				-0.21		-0.01					0.37					0.026			0.03	
phi_21				-0.04		-0.02					0.37					-0.05			0.13	
phi_22				0.01		-0.04					0.28					-0.16			0.08	
phi_23				0.07		-0.09					0.39					-0			0.03	
phi_24				-0.4		-0.44										-0.45			-0.5	
PHI_1	-0.2		-0.7	0.13	-0.73	-1.28	-0.7	-0.7	-0.7	-0.718	-1.2	-0.8	-0.8	-0.855	-0.8	-1.24	-1	-0.9	0.26	-0.618
PHI_2	-0.5		-0.5	-0.41	-0.57	-0.88	-0.5	-0.5	-0.5	-0.548	-0.6	-0.5	-0.5	-0.606	-0.6	-0.6	-0.6	-0.5	-0.1	-0.399
PHI_3	-0.3		-0.2	-0.13	-0.17	-0.19	-0.2	-0.3	-0.1	-0.14		-0.1	-0.1	-0.173	-0.2		-0.1	-0.1	-0.3	-0.134

En esta etapa de estimación de parámetros bajo el *Sieve Bootstrap* se considera el método de Mínimos Cuadrados Condicionales. Las tablas 5.9 y 5.10 muestran las estimaciones de los parámetros identificados bajo el AIC y AICC, respectivamente.

Tabla 5.10: Estimación de Parámetros de los Modelos Sieve Bootstrap con el AICC

COEFICIENTES AUTORREGRESIVOS REGULARES & ESTACIONALES CON SIEVE BOOTSTRAP ,AICC,B=200,1000,2000,h=1,6,12																				
Coef.	*SCA*	*SCTP*	*SCTT*	*SCOT*	*SCWP*	*SA*	*STP*	*STT*	*SOT*	*SWP*	*NSA*	*NSTP*	*NSTT*	*NSOT*	*NSWP*	*DCA*	*DCTP*	*DCTT*	*DCOT*	*DCWP*
Cte.	0	0	0	0	0	0	0	0	0	-0.003	0	0	0.002	-0.001	0	-0.02	0	0	0	0
phi_1	-0.62	-0.5	-0.6	-0.48	-0.5	0.1	-0.6	-0.5	-0.7	0.3072	-0.9	-0.72	-0.78	0.3187	0.299	0.179	-0.72	-0.72	-0.52	-0.68
phi_2	-0.41	-0.6	-0.55	-0.31	-0.33	0.3	-0.4	-0.4	-0.5	0.1513	-0.6	-0.52	-0.69	0.2374	0.257	0.003	-0.46	-0.53	-0.31	-0.47
phi_3	-0.39	-0.3	-0.23	-0.23	-0.24		-0.3	-0.3	-0.4		-0.5	-0.38	-0.58			-0.01	-0.31	-0.39	-0.14	-0.41
phi_4	-0.22	-0.3	-0.09	-0.19	-0.19		-0.2	-0.2	-0.4		-0.4	-0.24	-0.43			-0.03	-0.43	-0.48	-0.11	-0.31
phi_5	-0.11	-0.2					-0.2	-0.2	-0.3		-0.3	-0.17	-0.38			-0	-0.4	-0.43	-0.15	-0.31
phi_6	-0.13	-0.2						-0.1	-0.2		-0.2	-0.22	-0.39			-0.11	-0.25	-0.3	-0.21	-0.3
phi_7	-0.16	-0.1						-0.1	-0.2			-0.12	-0.32			0.038	-0.2	-0.26	-0.12	-0.23
phi_8	-0.26	-0.2							-0.1			-0.12	-0.31			0.067	-0.26	-0.29	-0.04	-0.16
phi_9	-0.24	-0.2							-0.2				-0.21			-0.05	-0.32	-0.3	0.056	
phi_10	-0.08	-0.1											-0.19			-0	-0.28	-0.3	0.029	
phi_11	-0.09	-0.1														-0.04	-0.19	-0.23	-0	
phi_12	-0.59	0.09														0.369		-0.14	-0.89	
phi_13	-0.3	0.07														-0.03			-0.34	
phi_14	-0.21	0.02														0.02			-0.25	
phi_15	-0.21	-0.1														0.034			-0.05	
phi_16																-0.09			-0.04	
phi_17																0.013			-0.11	
phi_18																0.042			-0.16	
phi_19																0.008			-0.06	
phi_20																0.026			0.03	
phi_21																-0.05			0.133	
phi_22																-0.16			0.079	
phi_23																-0			0.031	
phi_24																-0.45			-0.46	
PHI_1	-0.22		-0.66	-0.73	-0.73	-1	-0.7	-0.7	-0.7	-0.718	-0.7	-0.82	-0.76	-0.855	-0.83	-1.24	-1	-0.89	0.261	-0.62
PHI_2	-0.48		-0.52	-0.58	-0.57	-1	-0.5	-0.5	-0.5	-0.548	-0.3	-0.52	-0.5	-0.606	-0.6	-0.6	-0.59	-0.51	-0.15	-0.4
PHI_3	-0.31		-0.24	-0.16	-0.17	-0	-0.2	-0.3	-0.1	-0.14	-0.1	-0.13	-0.1	-0.173	-0.18		-0.15	-0.12	-0.34	-0.13

- **Diagnostico bajo el Sieve Bootstrap.**

En estudio de *Monte Carlo* para la aproximación *Sieve Bootstrap*, al conocer los residuales de los modelos identificados y estimados, estos residuales inicialmente son centrados, esto es que se calculan los residuales restados en su promedio, y a partir de ellos se construye la distribución empírica de estos, y luego el procedimiento de remuestreo obteniendo muestras *bootstrap* de los residuales.

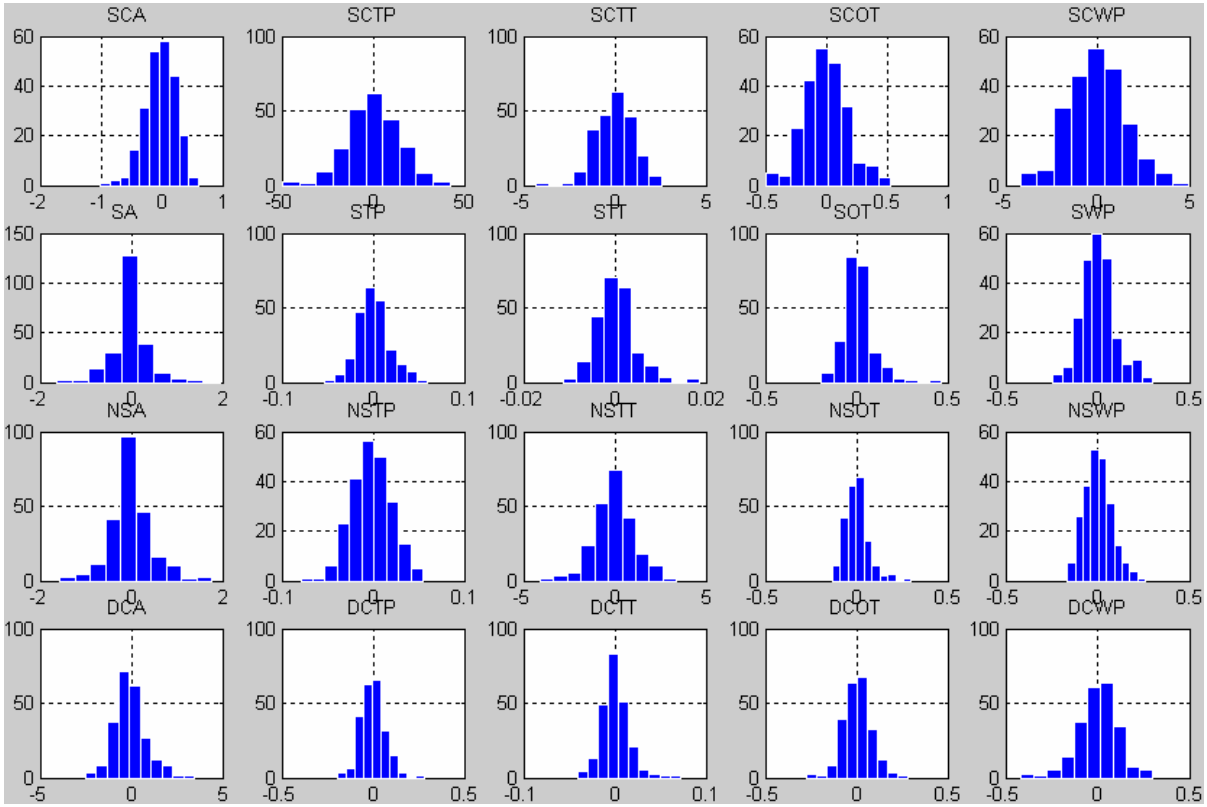


Figura 5.10: Histogramas de Residuos de los Modelos Seleccionados con el AIC, bajo el Sieve Bootstrap

En las figuras 5.10 y 5.11, se presentan el histograma y las funciones de autocorrelación de los residuales de los modelos seleccionados con el AIC.

Aparentemente las figuras presentan un buen diagnóstico de los modelos ajustados mediante el AIC, ya que los histogramas se comportan de forma simétrica, y las funciones de autocorrelación de los residuos nos sugieren que estos, se comportan como un ruido blanco.

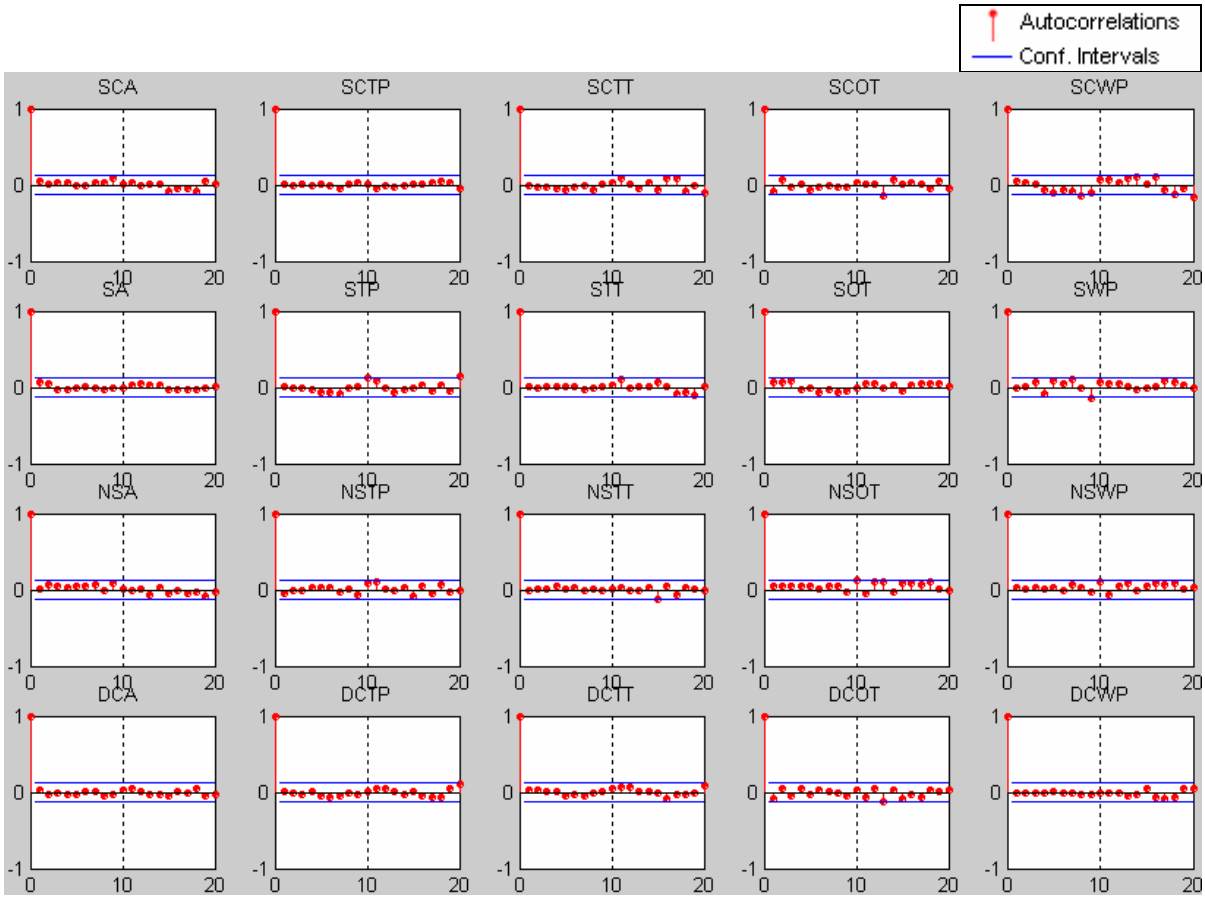


Figura 5.11: Funciones de Autocorrelación de Residuos de los Modelos Seleccionados con el AIC, bajo el Sieve Bootstrap

El comportamiento de los residuales de los modelos *Sieve Bootstrap* seleccionados con el AIC, mostró un comportamiento de ruido blanco, por otro lado en la tabla 5.11 se muestra la prueba de *Jarque Bera* para ver si la distribución de los residuales sigue una distribución normal.

Tabla 5.11: Prueba de Normalidad de Jarque Bera para los Modelos Sieve Bootstrap seleccionados mediante el AIC.

Series	Jarque Bera	P-valor
SCA	9.5647	8.40E-03
SCTP	2.0052	0.3669
SCTT	11.4793	3.20E-03
SCOT	1.2972	0.5228
SCWP	0.1066	0.9481
SA	204.9967	0
STP	3.8572	0.1454
STT	56.7034	4.86E-13
SOT	329.2754	0
SWP	12.7159	1.70E-03
NSA	41.6739	8.93E-10
NSTP	0.7348	0.6925
NSTT	4.6521	0.0977
NSOT	92.4723	0
NSWP	5.3567	6.87E-02
DCA	27.08	1.32E-06
DCTP	17.5297	1.56E-04
DCTT	70.9375	4.44E-16
DCOT	16.0428	3.28E-04
DCWP	26.1448	2.10E-06

Mediante la prueba de normalidad de *Jarque Bera* estadísticamente los residuales de las series SCTP, SCOT, SCWP, STP, NSTP, NSTT y NSWP. Siguen una distribución normal con un nivel de significación del 5%.

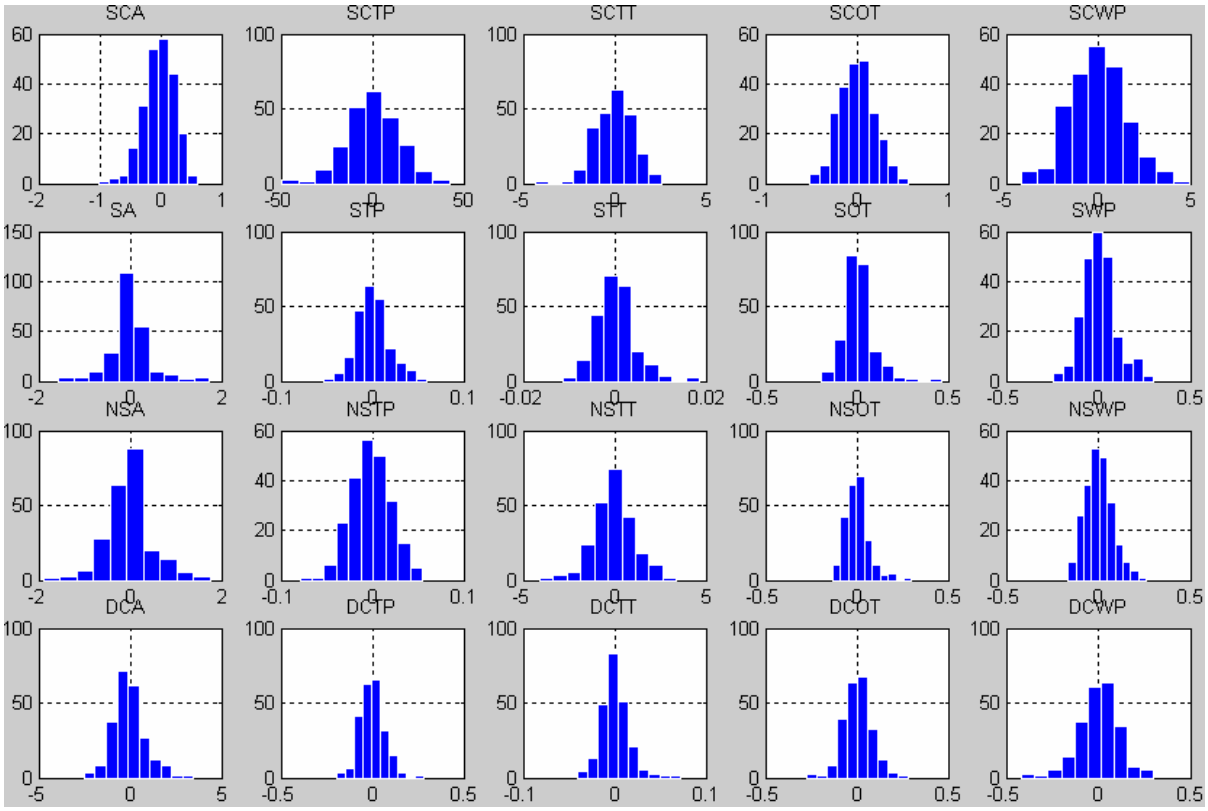


Figura 5.12: Histogramas de Residuos de los Modelos Seleccionados con el AICC, bajo el Sieve Bootstrap

Las figuras 5.12 y 5.13 presentan los histogramas y las funciones de autocorrelación de los residuos pertenecientes a los modelos seleccionados mediante el AICC, análogamente estos residuos presentan una forma simétrica en sus distribuciones y por otro lado las funciones de autocorrelación presentan que los residuos se comportan como ruido blanco.

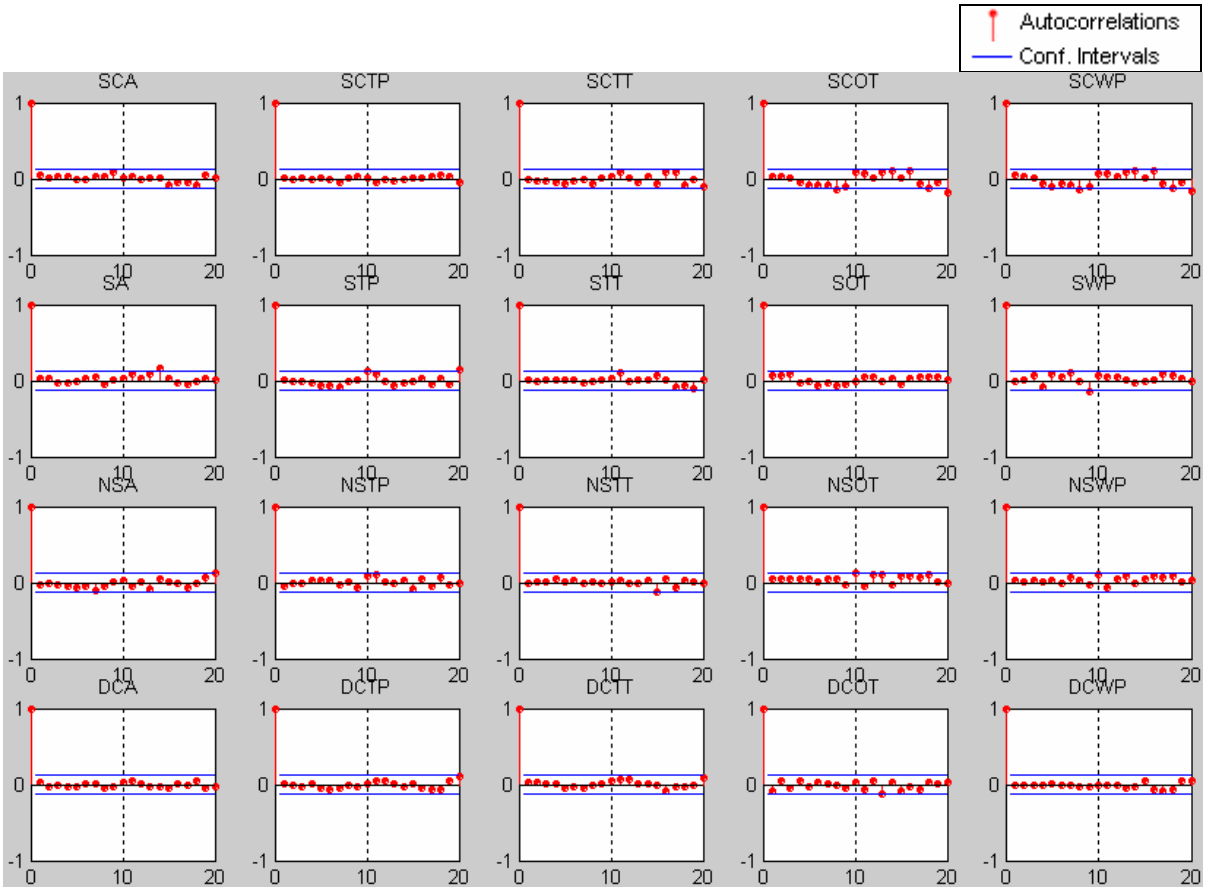


Figura 5.13: Funciones de Autocorrelación de los Residuos de los Modelos Seleccionados con el AICC, bajo el Sieve Bootstrap

El comportamiento de los residuales de los modelos *Sieve Bootstrap* seleccionados mediante el AICC, análogamente mostró un comportamiento de ruido blanco, por otro lado en la tabla 5.12 se muestra la prueba de *Jarque Bera* para ver si la distribución de los residuales sigue una distribución normal.

Tabla 5.12: Prueba de Normalidad de Jarque Bera para los Modelos Sieve Bootstrap seleccionados mediante el AICC.

Series	Jarque Bera	P-valor
SCA	9.5647	8.40E-03
SCTP	2.0052	0.3669
SCTT	11.4793	3.20E-03
SCOT	0.0371	0.9816
SCWP	0.1066	0.9481
SA	114.9431	0
STP	3.8572	0.1454
STT	56.7034	4.86E-13
SOT	329.2754	0
SWP	12.7159	1.70E-03
NSA	36.6308	1.11E-08
NSTP	0.7348	0.6925
NSTT	4.6521	0.0977
NSOT	92.4723	0
NSWP	5.3567	6.87E-02
DCA	27.08	1.32E-06
DCTP	17.5297	1.56E-04
DCTT	70.9375	4.44E-16
DCOT	16.0428	3.28E-04
DCWP	26.1448	2.10E-06

Mediante la prueba de normalidad de *Jarque Bera* estadísticamente los residuales de las series SCTP, SCOT, SCWP, STP, NSTP, NSTT y NSWP. Siguen una distribución normal con un nivel de significación del 5%, cuyos resultados son análogos a los de los modelos *Sieve Bootstrap* mediante el AIC.

La aproximación *Sieve Bootstrap* mostrada para las 20 series de tiempo representan modelos adecuados para describir los datos, consecuentemente se realiza una validación para los modelos propuestos en la tabla 5.7 y 5.8, tomando como muestra de entrenamiento $n = 246$, que comprende el periodo Julio/1983 hasta Diciembre/2003; y como muestra de prueba

$n = 12$ que comprende el periodo Enero/2004 hasta Diciembre/2004. Por otro lado se consideran diferentes números de muestras *bootstrap* $B = 200, 1000, 2000$ con el mismo tamaño, para visualizar el efecto del número de iteraciones en el estudio Monte Carlo, de esta forma tenemos seis posibles combinaciones para los métodos de selección y la cantidad de muestras *bootstraps.*

- **Validación del Sieve Bootstrap bajo el AIC, B=200 y h=12.**

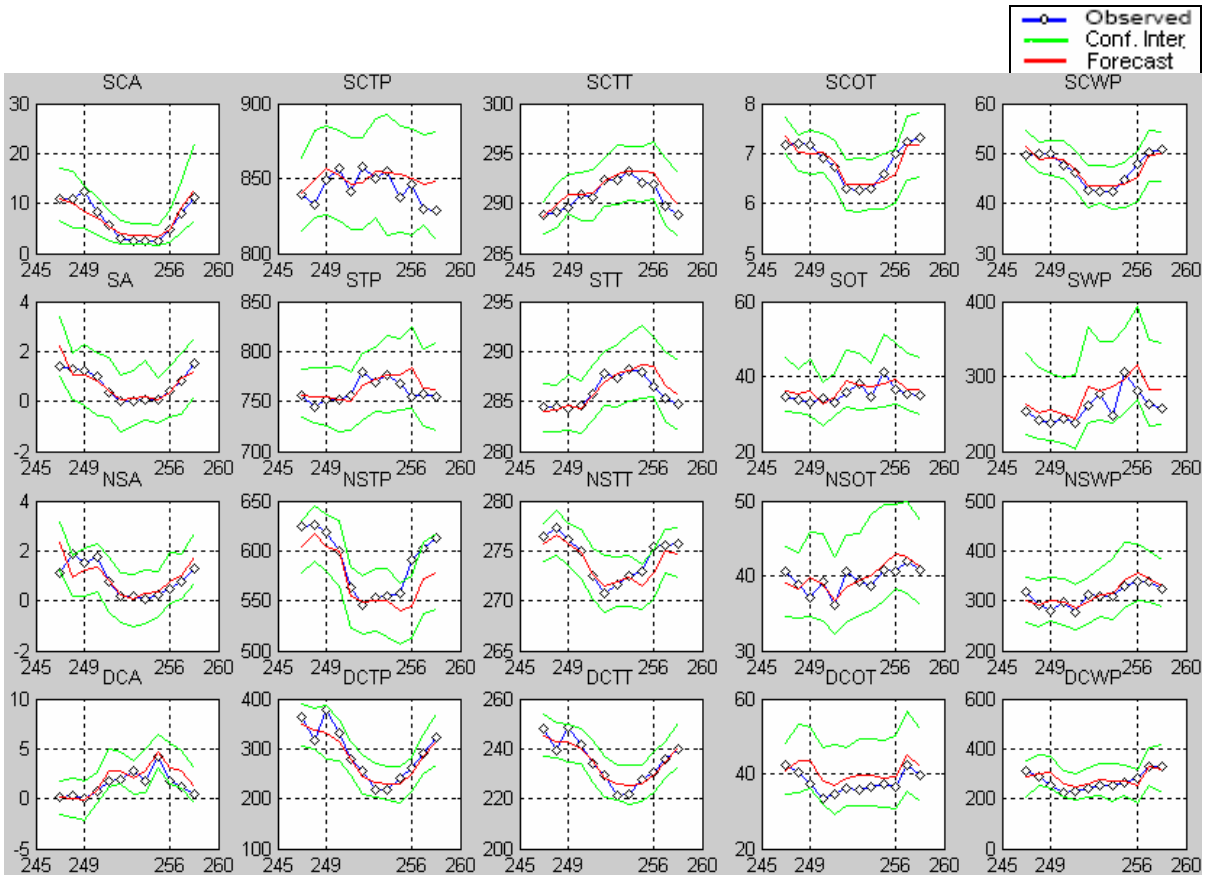


Figura 5.14: Validación para el Sieve Bootstrap bajo AIC, B=200 y h=12.

La figura 5.14, presenta la validación para el *sieve bootstrap* con la selección de modelos bajo el AIC, con 200 remuestras *bootstraps* de los residuales, dejando el año 2004 para la validación

- **Validación del Sieve Bootstrap bajo el AIC, B=1000 y h=12.**

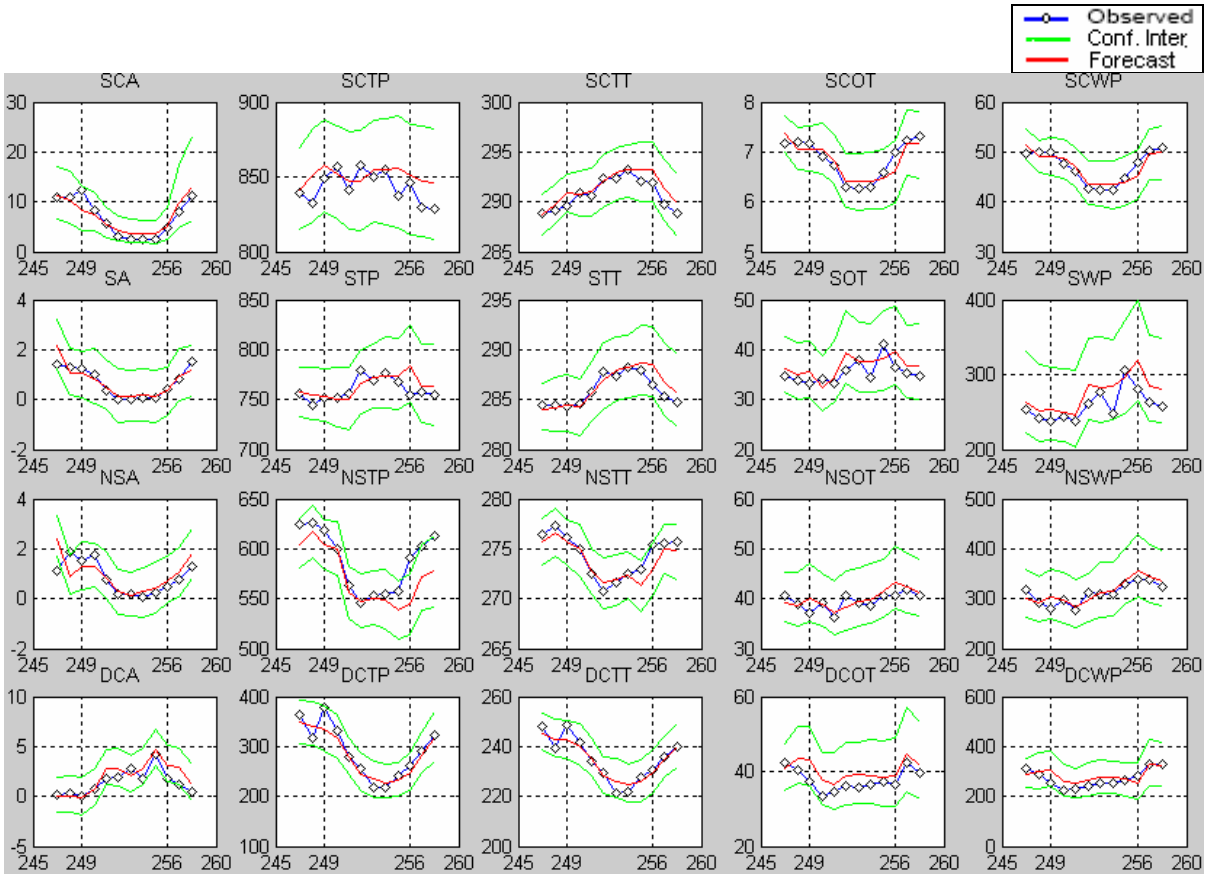


Figura 5.15: Validación para el Sieve Bootstrap bajo AIC, B=1000 y h=12.

La figura 5.15, presenta la validación para el *sieve bootstrap* con la selección de modelos bajo el AIC, con 1000 muestras *bootstrap* de los residuales, dejando el año 2004 para la validación.

- **Validación del Sieve Bootstrap bajo el AIC, B=2000 y h=12.**

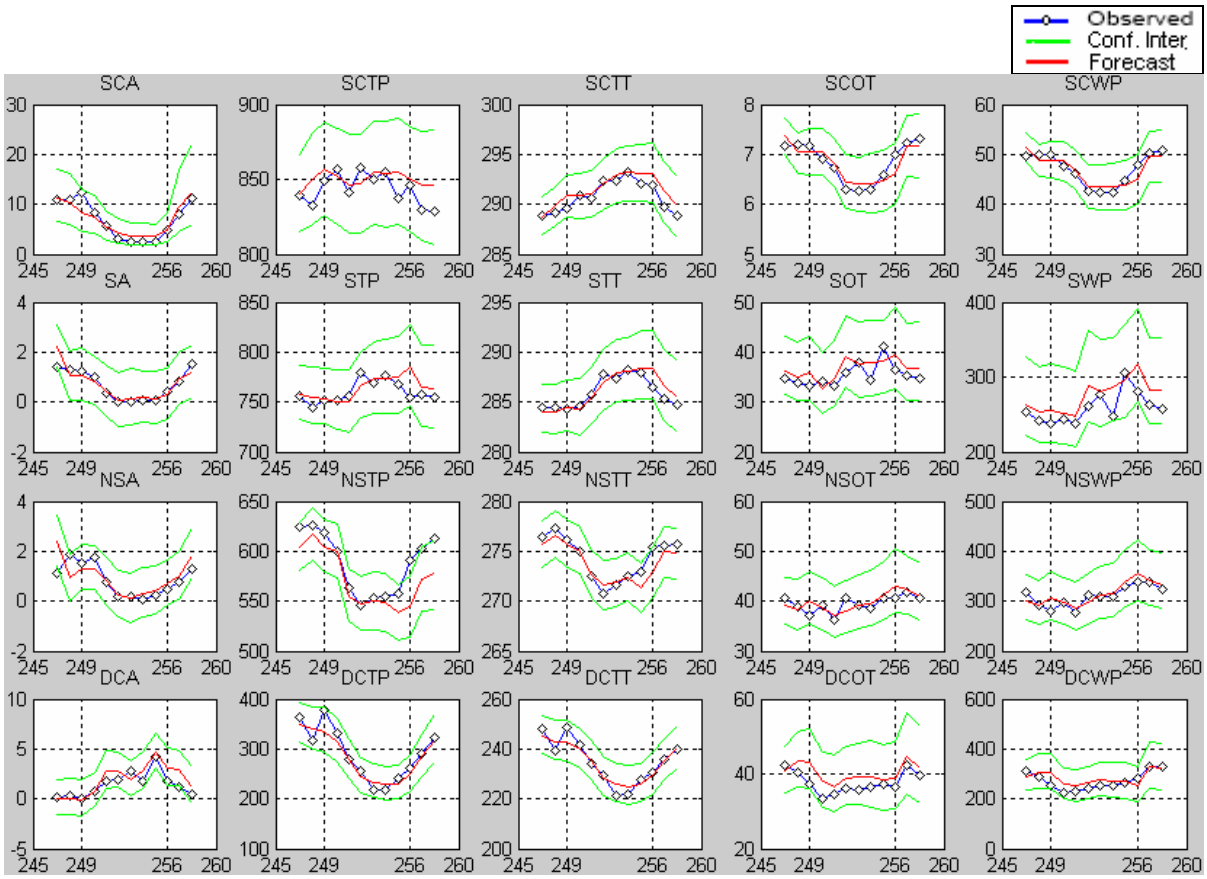


Figura 5.16: Validación para el Sieve Bootstrap bajo AIC, B=2000 y h=12.

La figura 5.16, presenta la validación para el *sieve bootstrap* con la selección de modelos bajo el AIC, con 2000 muestras *bootstrap* de los residuales, dejando el año 2004 para la validación.

- **Validación del Sieve Bootstrap bajo el AICC, B=200 y h=12.**

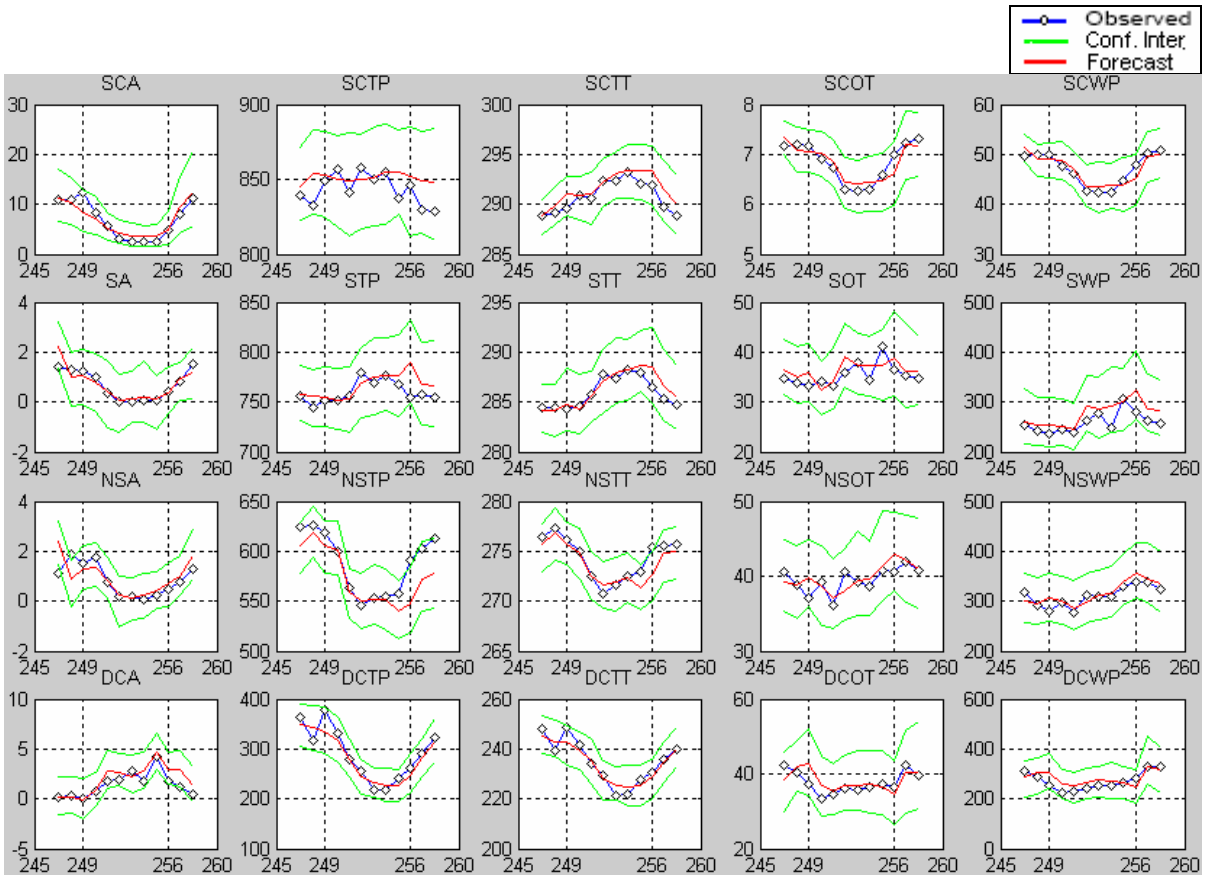


Figura 5.17: Validación para el Sieve Bootstrap bajo AICC, B=200 y h=12.

La figura 5.17, presenta la validación para el *sieve bootstrap* con la selección de modelos bajo el AICC, con 200 muestras *bootstrap* de los residuales, dejando el año 2004 para la validación.

- **Validación del Sieve Bootstrap bajo el AICC, B=1000 y h=12.**

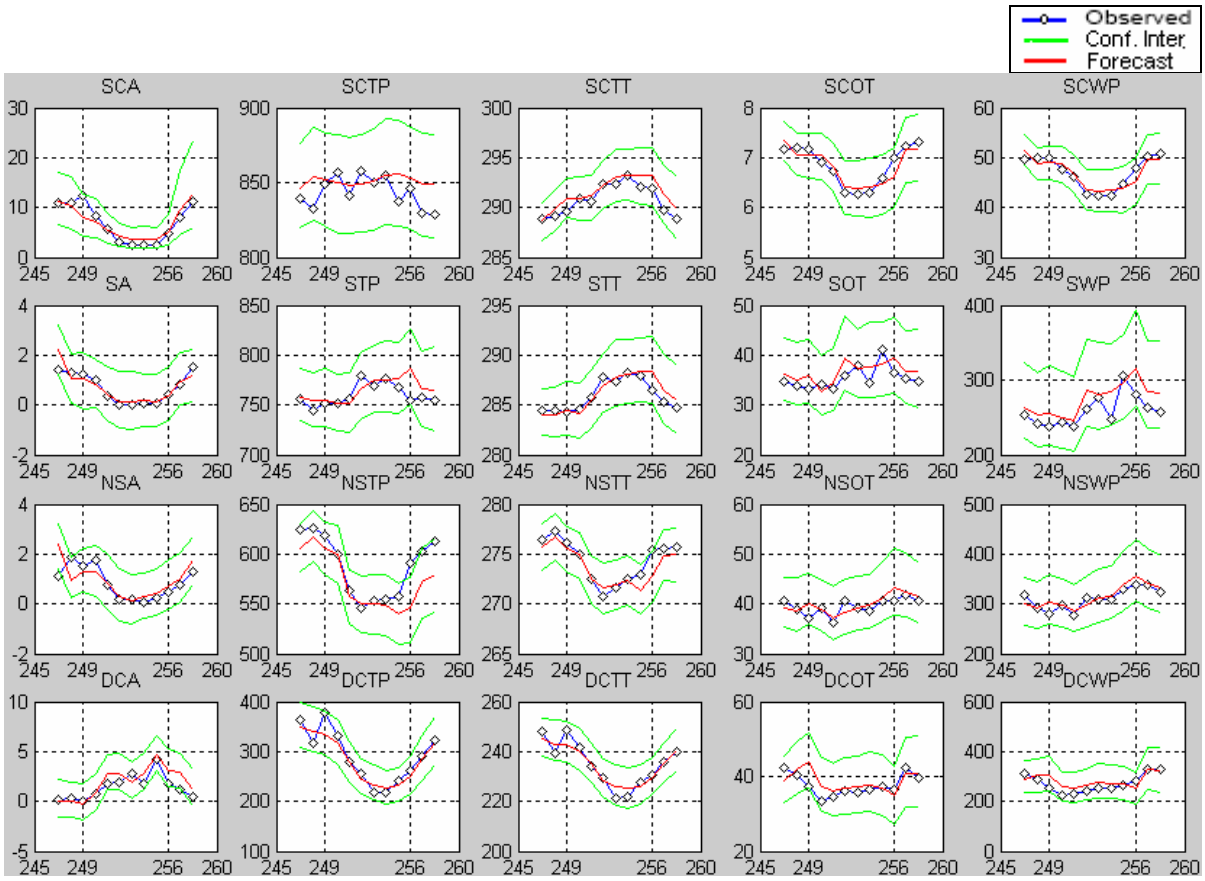


Figura 5.18: Validación para el Sieve Bootstrap bajo AICC, B=1000 y h=12.

La figura 5.18, presenta la validación para el *sieve bootstrap* con la selección de modelos bajo el AICC, con 1000 muestras *bootstrap* de los residuales, dejando el año 2004 para la validación.

- **Validación del Sieve Bootstrap bajo el AICC, B=2000 y h=12.**

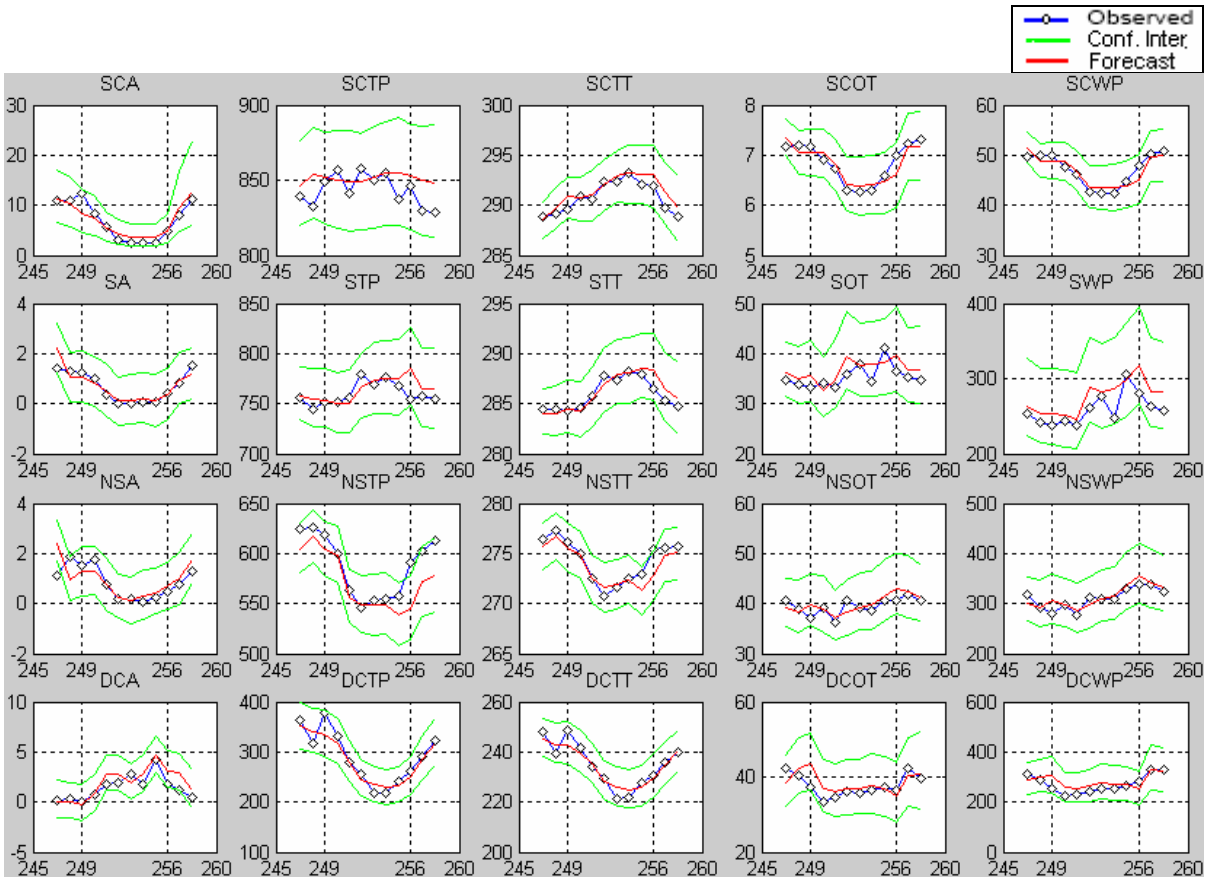


Figura 5.19: Validación para el Sieve Bootstrap bajo AICC, B=2000 y h=12.

La figura 5.19, presenta la validación para el *sieve bootstrap* con la selección de modelos bajo el AICC, con 2000 muestras *bootstrap* de los residuales, dejando el año 2004 para la validación.

Finalmente se considera las predicciones para el año 2005, para los modelos seleccionados con el AIC y AICC, cunado el número de muestras *bootstrap* en los residuales, consecuentemente en las predicciones sean B = 1000.

- **Predicciones para el año 2005 bajo el Sieve Bootstrap con AIC, B=1000, h=12.**

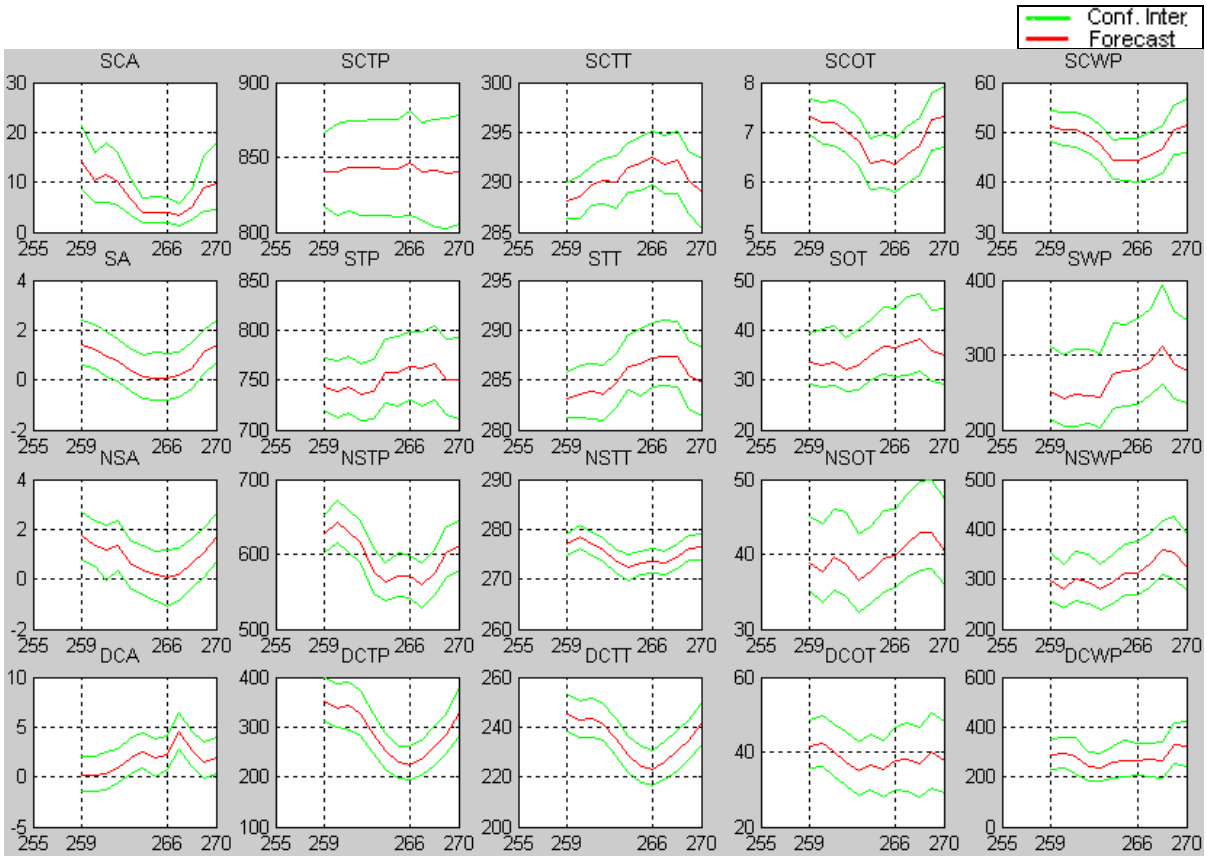


Figura 5.20: Predicciones para el Sieve Bootstrap con AIC, B=1000 para Enero/2005 a Diciembre/2005.

La figura 5.20, presenta las predicciones e intervalos de predicción con un 95% de confianza, para el año 2005 utilizando el *Sieve Bootstrap* con el criterio de elección AIC, con el número de muestras *bootstrap* B=1000 y h=12.

- **Predicciones para el año 2005 bajo el Sieve Bootstrap con AICC, B=1000, h=12.**

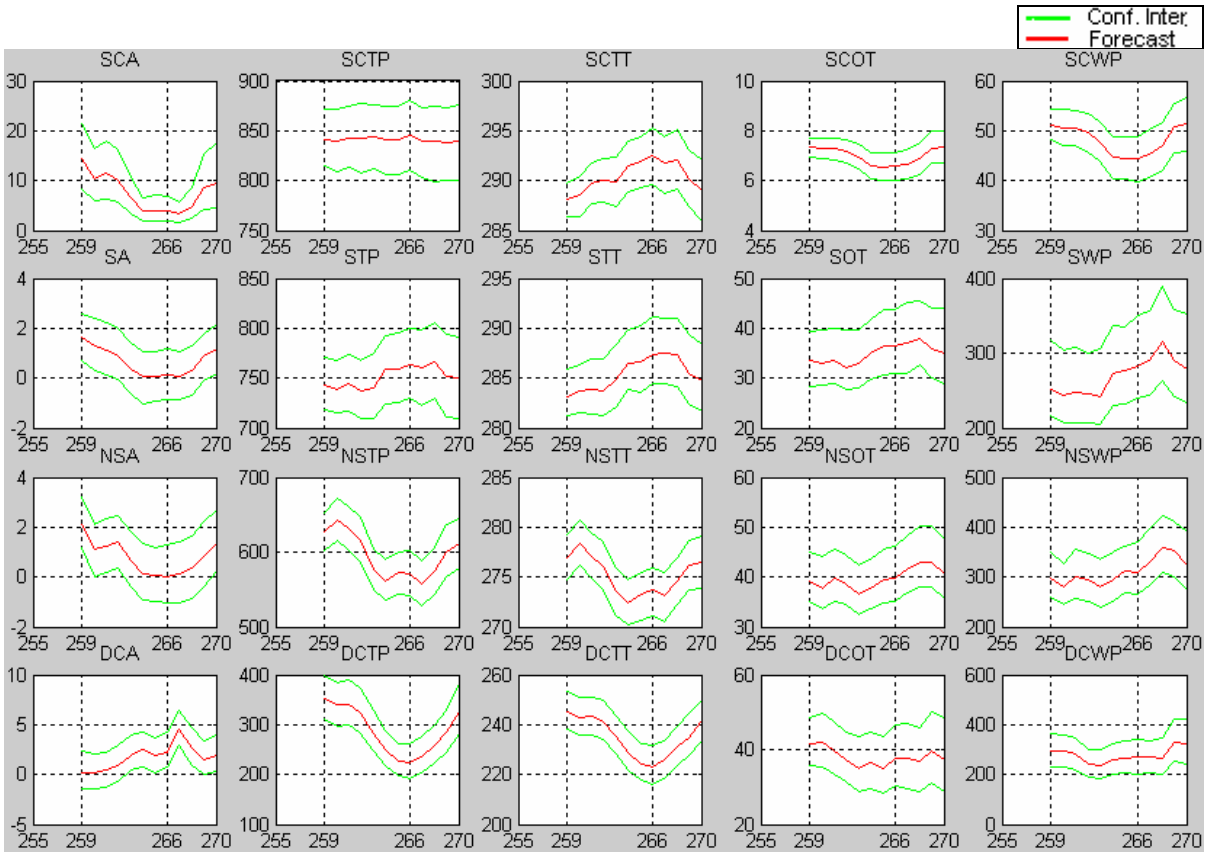


Figura 5.21: Predicciones para el Sieve Bootstrap con AICC, B=1000 para Enero/2005 a Diciembre/2005.

La figura 5.21, presenta las predicciones e intervalos de predicción con un 95% de confianza, para el año 2005 utilizando el *Sieve Bootstrap* con el criterio de elección AICC, con el número de muestras *bootstrap* B=1000 y h=12. Análogamente a las predicaciones de los modelos *Box y Jenkins*, estas siguen el mismo patrón estacional de las series de nubosidad.

5.3 Precisión de un Intervalo de Predicción

La precisión de cada intervalo de predicción, se estima usando el porcentaje de cobertura de las observaciones, dado por: $C_M = \#\{Q_M(\alpha/2) \le X_{T+h} \le Q_M(1-\alpha/2)\}/h$, simultáneamente se puede calcular la amplitud del intervalo de predicción, usando $L_M = Q_M(1-\alpha/2) - Q_M(1-\alpha/2)$, donde M representa el método utilizado. Para nuestro conjunto de datos se evaluara la precisión y amplitud de cada intervalo de predicción para los métodos *Box Jenkins* y *Sieve Bootstrap*, y así poder comprar sus respectivas eficiencias. El procedimiento de comparación se realiza dejando el año 2004 para la respectiva validación.

Por otro lado los modelos *Box Jenkins* y *Sieve Bootstrap* son comparables por que ambos tienen una diagnosis correcta, es decir, ambos dejan sin estructura los residuos, así dos modelos distintos pueden ser satisfactorios para una misma serie en términos de diagnosis. Ya comprobado esto, la comparación de los dos modelos es en cuanto a la predicción consecuentemente sus intervalos de predicción. Queda destacar que para cada modelo considerado en el *Sieve Bootstrap* en todos los casos se centraron los residuales, teniendo así media cero y varianza constante.

5.4 Resultados

Inicialmente se comparan los métodos de *Box Jenkins* y *Sieve Bootstrap*, mediante sus intervalos de predicción, esta comparación es realizada mediante una validación en los dos métodos, dejando el año 2004 para la validación, de esta forma calcular las medidas de precisión de los intervalos de predicción y compararlas.

Los resultados obtenidos con los métodos *Box Jenkins* y *Sieve Bootstrap*, se encuentran en la tabla 5.13 y 5.14, los valores en negrita identifican que un Intervalo de predicción obtuvo una mayor cobertura y una menor amplitud respectivamente.

Comparación de Coberturas de los Intervalos de Predicción

Las comparaciones de cobertura de los modelos *Box Jenkins* y *Sieve Bootstrap* con AIC, AICC, a diferentes B = 200, 1000, 2000; para las últimas doce observaciones, sin embargo únicamente se muestran las figuras de las combinaciones AIC con B = 1000 y AICC con B= 1000, los resultados se presentan a continuación:

- **Comparación de Cobertura de los Modelos Box Jenkins y Sieve Bootstrap bajo el AIC, B=1000 y h=12.**

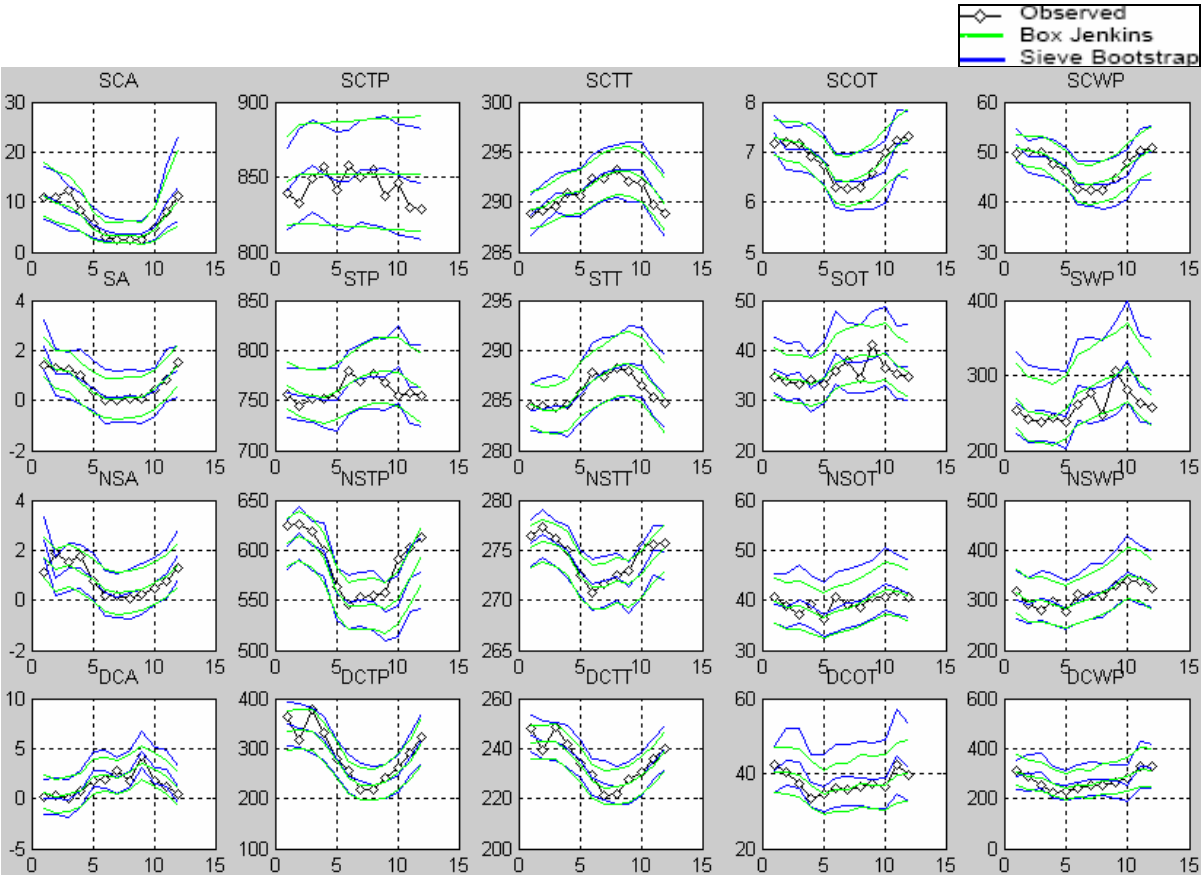


Figura 5.22: Predicciones e Intervalos de Predicción al 95% de Box Jenkins y Sieve Bootstrap bajo AIC, B=1000 y h=12.

La figura 5.22, presenta la comparación de las predicciones e intervalos de predicción con un nivel de significación del 5%, de los modelos *Box Jenkins* estimados en la tabla 5.4, y los modelos *Sieve Bootstrap* con AIC, B=1000, ambos validando el año 2004, para las series de tiempo de nubosidad.

- **Comparación de Cobertura de los Modelos Box Jenkins y Sieve Bootstrap bajo el AICC, B=1000 y h=12.**

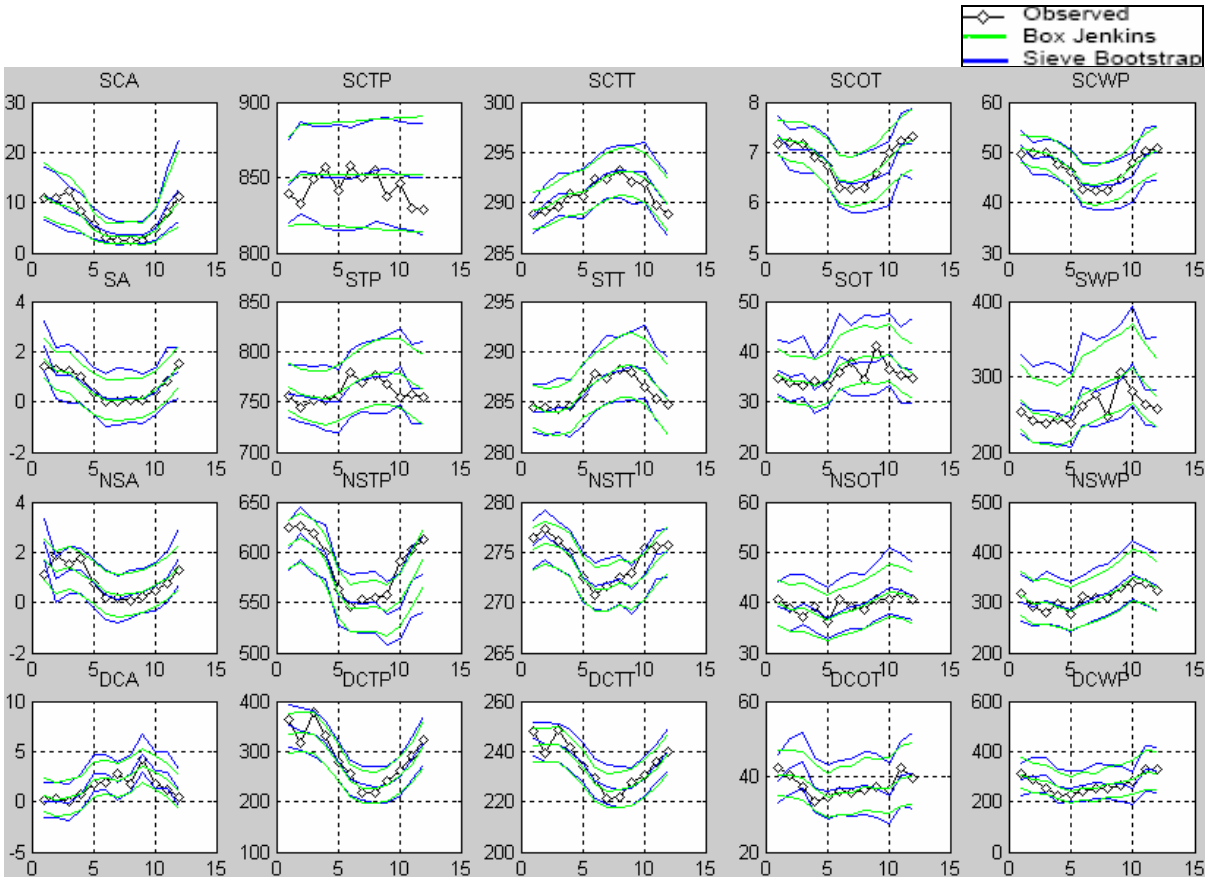


Figura 5.23: Predicciones e Intervalos de Predicción al 95% de Box Jenkins y Sieve Bootstrap bajo AICC, B=1000 y h=12.

La figura 5.23, presenta la comparación de las predicciones e intervalos de predicción con un nivel de significación del 5%, de los modelos *Box Jenkins* estimados en la tabla 5.4, y los modelos *Sieve Bootstrap* con AICC, B=1000, ambos validando el año 2004, para las series de tiempo de nubosidad.

A continuación se muestran los resultados tabularmente:

Tabla 5.13: Cobertura de Intervalos de Predicción de 95% de los Modelos Box Jenkins y Sieve Bootstrap con AIC, AICC, B = 200, 1000, 2000 y h = 12

PORCENTAJE DE COBERTURAS DE INTERVALOS DE PREDICCION AL 95% DE LOS MODELOS BOX JENKINS Y SEIVE BOOTSTRAP, CON h=12.							
		AIC			*AICC*		
Series	*Box Jenkins*	*B=200*	***B=1000***	*B=2000*	*B=200*	*B=1000*	*B=2000*
SCA	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SCTP	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SCTT	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SCOT	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SCWP	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SA	100.0	100.0	100.0	100.0	100.0	100.0	100.0
STP	100.0	100.0	100.0	100.0	100.0	100.0	100.0
STT	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SOT	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SWP	91.67	100.0	100.0	100.0	100.0	100.0	100.0
NSA	100.0	91.67	91.67	91.67	91.67	91.67	91.67
NSTP	83.33	91.67	91.67	91.67	91.67	91.67	91.67
NSTT	91.67	91.67	100.0	91.67	91.67	91.67	91.67
NSOT	100.0	100.0	100.0	100.0	100.0	100.0	100.0
NSWP	100.0	100.0	100.0	100.0	100.0	100.0	100.0
DCA	100.0	100.0	100.0	100.0	100.0	100.0	100.0
DCTP	91.67	100.0	100.0	100.0	91.67	100.0	100.0
DCTT	100.0	100.0	100.0	100.0	100.0	100.0	100.0
DCOT	100.0	100.0	100.0	100.0	100.0	100.0	100.0
DCWP	100.0	100.0	100.0	100.0	100.0	100.0	100.0

La tabla 5.13 presenta la comparación de las coberturas de los intervalos de predicción al 95% de confianza, observando los resultados la cobertura de los modelos ajustados mediante la metodología *Box Jenkins* no presenta una buena cobertura, por otro lado la cobertura de los intervalos ajustados mediante el *Sieve Bootstrap* ofrecen una mejor cobertura, de esta manera, la mejor combinación de criterios y numero de muestras bootstraps, es la de AIC con B = 1000.

Comparación de Amplitudes de los Intervalos de Predicción

Las comparaciones de amplitudes de los modelos *Box Jenkins* y *Sieve Bootstrap* con AIC, AICC, a diferentes B = 200, 1000, 2000; para las últimas doce observaciones, análogamente se presentan únicamente las figuras cuyas combinaciones son AIC con B = 1000 y AICC con B = 1000, los resultados son como siguen:

- **Comparación de Amplitud de los Modelos Box Jenkins y Sieve Bootstrap bajo el AIC, B=1000 y h=12.**

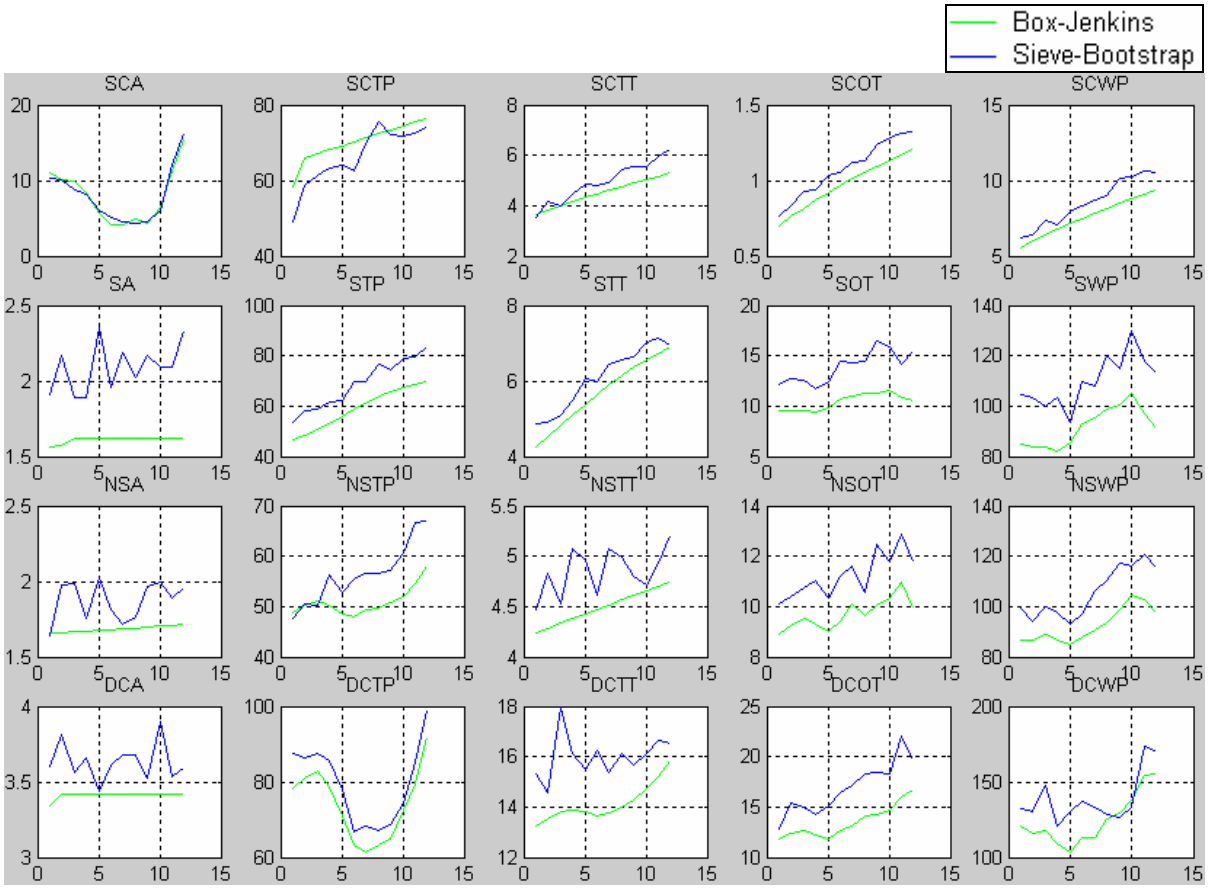


Figura 5.24: Amplitudes de Intervalos de Predicción al 95% de Box Jenkins y Sieve Bootstrap con AIC, B=1000 y h=12.

La figura 5.24, presenta las amplitudes de los intervalos de predicción con un nivel de significación del 5%, de los modelos *Box Jenkins* estimados en la tabla 5.4, y los modelos *Sieve Bootstrap* con AIC, B=1000, ambos validando el año 2004, para las series de tiempo de nubosidad.

- **Comparación de Amplitud de los Modelos Box Jenkins y Sieve Bootstrap bajo el AICC, B=1000 y h=12.**

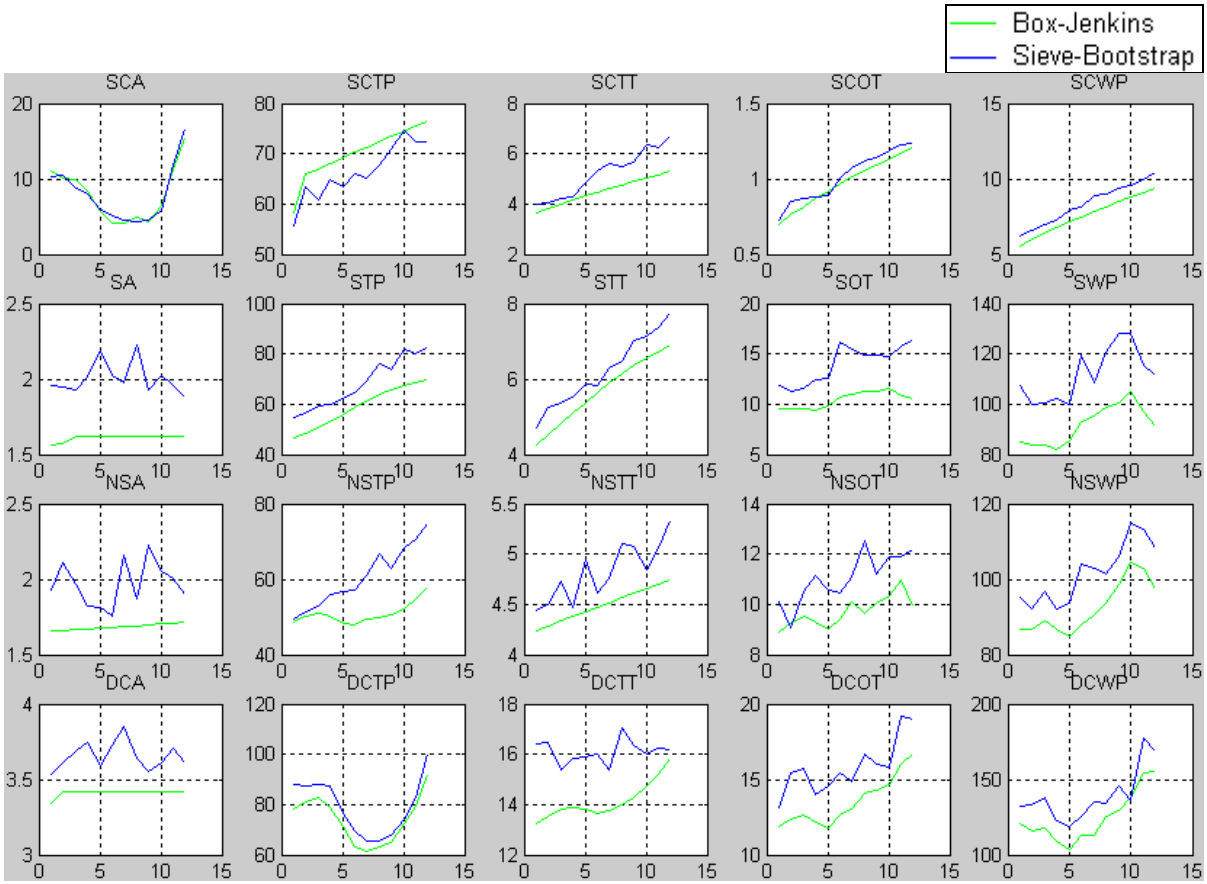


Figura 5.25: Amplitudes de Intervalos de Predicción al 95% de Box Jenkins y Sieve Bootstrap con AICC, B=1000 y h=12.

La figura 5.25, presenta las amplitudes de los intervalos de predicción con un nivel de significación del 5%, de los modelos *Box Jenkins* estimados en la tabla 5.4, y los modelos *Sieve Bootstrap* con AICC, B=1000, ambos validando el año 2004, para las series de tiempo de nubosidad.

Tabla 5.14: Amplitud de Intervalos de Predicción de los Modelos Box Jenkins y Sieve Bootstrap con AIC, AICC, B = 200, 1000, 2000 y h = 12.

PROMEDIOS DE AMPLITUDES DE INTERVALOS DE PREDICCION AL 95% DE LOS MODELOS BOX JENKINS Y SEIVE BOOTSTRAP, CON h=12.							
Series	*Box Jenkins*	*AIC*			*AICC*		
		B=200	*B=1000*	*B=2000*	*B=200*	*B=1000*	*B=2000*
SCA	7.9834	8.2676	7.9421	7.9138	**7.5962**	8.1164	8.0425
SCTP	70.169	67.078	66.012	66.805	**63.335**	66.152	66.524
SCTT	**4.5108**	5.15	5.0913	5.1305	4.9552	4.9435	5.1353
SCOT	**0.97534**	1.0735	1.0651	1.0557	1.0391	1.0634	1.0896
SCWP	**7.5735**	8.2782	8.6432	8.5472	8.5173	8.1941	8.494
SA	**1.6104**	2.0525	2.0698	2.033	2.1363	2.1014	2.0092
STP	**59.179**	66.856	67.211	68.474	71.086	66.279	67.741
STT	**5.6857**	6.221	6.0683	6.071	6.1288	5.9866	6.0593
SOT	**10.429**	13.129	13.129	13.995	12.959	13.854	13.944
SWP	**91.706**	107.72	110.7	111.59	112.58	110	110.91
NSA	**1.6842**	1.8676	1.9253	1.8979	1.7532	1.9189	1.8905
NSTP	**50.938**	54.567	58.337	56.648	58.141	59.443	59.494
NSTT	**4.4923**	4.8528	4.7911	4.8309	4.922	4.929	4.8611
NSOT	**9.6969**	10.822	10.977	10.873	10.43	11.222	11.097
NSWP	**92.441**	105.46	101.67	103.21	102.58	105.31	103.92
DCA	**3.409**	3.562	3.6599	3.6051	3.5955	3.6216	3.6075
DCTP	**74.031**	79.183	82.717	79.26	77.308	81.569	82.593
DCTT	**14.136**	16.031	15.68	15.982	15.62	16.032	16.039
DCOT	**13.528**	16.233	16.602	16.782	16.963	15.486	15.553
DCWP	**124.66**	140.48	138.67	142.88	140.48	137.31	138.12

En la tabla 5.14, se muestran las amplitudes promedios de los intervalos de predicción al 95% de confianza de los modelos *Box Jenkins,* y *Sieve Bootstrap* con el criterio de elección AIC y AICC, y con un número de muestras *bootstrap* de B = 200, 1000 y 2000, y finalmente

con un periodo de predicción h =12. Esta comparación hecha en la tabla 5.14 muestra que la amplitud de los intervalos de predicción calculados mediante el método de *Box Jenkins* en la mayoría son menores en magnitud comparadas con las amplitudes de los intervalos de predicción bajo el Sieve Bootstrap, no obstante entre las amplitudes de los intervalos de predicción construidos mediante el método del *Sieve Bootstrap* para diferentes criterios de elección, número de muestras *bootstrap*, y un periodo de predicción de doce unidades adelante, el que mejor amplitud tiene es la combinación del criterio AICC y B = 200 además que las amplitudes bajo esta combinación no difieren demasiado de los *Box Jenkins*.

En la tabla 5.14, para probar estadísticamente si existe diferencia significativa entre los métodos de *Box Jenkins* y *Sieve Bootstrap*, en las amplitudes de los intervalos de predicción, se empleó la prueba no paramétrica de *Wilcoxon*, debido a que las diferencias no siguen una distribución normal, caso contrario se pudo haber usado la prueba de t-student pareada. La prueba de *Wilcoxon* está basada en la diferencia de dos muestras pareadas, estas diferencias están dadas para la el método *Box Jenkins* y las combinaciones del *Sieve Bootstrap*, existiendo así seis diferencias, los resultados son:

Test de Wilcoxon para Amplitudes

	N Test	Wilcoxon Statistic	P-valor	Estimated Median
D1	20	14	0.001	-1.936
D2	20	15	0.001	-2.423
D3	20	14	0.001	-2.68
D4	20	19	0.001	-1.917
D5	20	14	0.001	-2.568

Como los P-valores son todos menores que 5%, entonces estadísticamente las amplitudes de los intervalos de predicción bajo los métodos de *Box Jenkins* y *Sieve Bootstrap* son distintas con 5% nivel de significación.

Finalmente, debido a que recientemente se obtuvieron los datos de las series de tiempo para el periodo enero a junio del 2005, se realiza la validación para dicho periodo utilizando los modelos Sieve Bootstrap mediante el criterio de selección *AIC y AICC*, con *B=1000* y *h=6.* El error absoluto de validación se muestra a continuación.

Sieve Bootstrap AIC

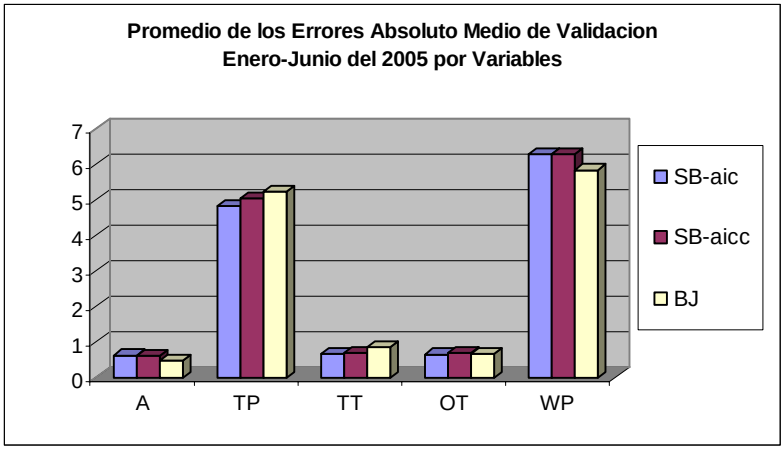
Error Absoluto Medio de Validación					
	A	TP	TT	OT	WP
SC	1.9044	5.4398	0.2725	0.1299	1.4365
S	0.166	4.54	1.0641	1.1756	7.4341
NS	0.2525	6.0952	0.4173	0.6081	3.0457
DC	0.1879	3.1961	0.9045	0.7123	13.227
mean	0.6277	4.817775	0.6646	0.656475	6.285825

Sieve Bootstrap AICC

Error Absoluto Medio de Validación					
	A	TP	TT	OT	WP
SC	1.9074	5.718	0.2772	0.203	1.3806
S	0.0677	4.7377	1.0534	1.1026	7.4335
NS	0.3193	6.1628	0.4391	0.634	2.9647
DC	0.1913	3.5488	0.9986	0.8	13.356
mean	0.621425	5.041825	0.692075	0.6849	6.2837

Box Jenkins

Error Absoluto Medio de Validación					
	A	TP	TT	OT	WP
SC	1.3827	5.3323	0.29343	0.15885	1.1175
S	0.075408	3.8861	1.1616	0.69341	7.6761
NS	0.19139	7.8742	0.63957	1.0116	5.897
DC	0.26047	3.7747	1.289	0.81919	8.6453
mean	0.477492	5.216825	0.8459	0.670763	5.833975



La comparación de los promedios de los errores absolutos medios de validación por variables en el periodo enero a junio del 2005, se ilustra en la grafica anterior, la cual exhibe un menor error absoluto medio de validación para el método de *Box Jenkins* en la variable *Amount y Water Path*, consecuentemente éste error es menor para el método *Sieve Bootstrap* mediante el *AIC* en la variable *Top Pressure* y *Top Temperatura.*

6 CONCLUSIONES Y TRABAJO FUTURO

6.1 Conclusiones

El *Sieve Bootstrap* en la construcción de intervalos de predicción para una clase general de modelos lineales incluyendo un *ARIMA* estacional, es una herramienta alternativa de buen desempeño, cuando se presenta que la distribución de los errores es normal ó aun cuando tiene una distribución asimétrica ya sea negativa o positiva. Este procedimiento de predicción remuestreando los residuales, no requiere especificar un modelo finito dimensional ni asumir distribuciones gausianas. Así, se ilustró que los intervalos de predicción construidos bajo la metodología *Box Jenkins* no tienen una precisión muy marcada cuando se recae en la suposición de normalidad, es decir que los intervalos de predicción están afectados por la suposición de normalidad. En cambio los intervalos de predicción bajo el *Sieve Bootstrap* muestran un mejoramiento marcado en la precisión de estos, es así que para el criterio de selección AIC y para el número de muestras bootstrap igual a 1000, se obtienen una cobertura del 100% casi en todas las series de tiempo.

Por otro lado al comparar las amplitudes de los intervalos de predicción con el fin de ver la variabilidad aparente en estos, los modelos *Box-Jenkins* presentan una mínima amplitud, comparados con el *Sieve Bootstrap* bajo los distintos criterios de elección y distintos números de muestras bootstrap, sin embargo bajo el *Sieve Bootstrap* la amplitud de los intervalos de

predicción comparativamente no distan mucho que la amplitud de los intervalos de predicción bajo el *Box-Jenkins*.

Finalmente las dos principales contribuciones de este trabajo, se centran en el estudio preliminar de los patrones de las nubes en el caribe, y un estudio analítico y metodológico de estos patrones, donde se mostró el buen desempeño de la técnica del *Sieve Bootstrap* para series de tiempo climatologicas.

6.2 Trabajo Futuro

En el análisis de series de tiempo surgen extensiones tanto para modelos lineales univariados como para modelos lineales multivariados, específicamente una trabajo futuro a considerar es la extensión del *Sieve Bootstrap* para la construcción de intervalos de predicción no parametricos y semi parametricos, en modelos de Función de Transferencia o modelos de Regresión Dinámica, que consideran una relación unidireccional entre dos ó más series de tiempo, una representación general de una modelo dinámico entre dos series de tiempo estacionarias es de la forma:

$$y_t = v_0 x_t + v_1 x_{t-1} + v_2 x_{t-2} + \ldots + n_t \quad \text{ó} \quad y_t = \beta_0 + \frac{w_0 + w_1 B}{(1 - \delta B)} B^b x_t + n_t$$

Donde y_t, x_t representan las dos series de tiempo, y η_t es un proceso que recoge la información de las otras variables, que generalmente sigue un proceso *ARMA(p,q)*.

En estos tipos de modelos donde la correspondencia es de causa efecto, se pueden también realizar extensiones simplistas y análogas a la presentada en este trabajo, tendiendo en cuenta la estructura de dependencia tanto univariada como bivariada. En el caso del estudio en cambios climatológicos en el caribe o en otras zonas, es importante considerar una correspondencia de causa efecto en algunas de las principales variables climatologicas y meteorológicas, como en impacto que tienen las nubes, los vientos, la humedad de suelo, la temperatura, en la precipitación de lluvia. Por este motivo el trabajo a futuro a considerar es modelo de función de transferencia no paramétrico para la construcción de intervalos de predicción.

REFERENCIAS

[1] Alonso, A. M. (2004). A Fortran Routine for sieve bootstrap prediction intervals. *Journal of Modern Applied Statistical Methods*, 3 (1), en prensa.

[2] Alonso, A. M., Peña, D., y Romo, J. (2003). On Sieve bootstrap prediction intervals. *Statistics & Probability Letters*, Elsevier, 65, 13-20.

[3] Alonso, A. M., Peña, D., y Romo, J. (2002). Una revisión de los métodos de Remuestreo en series temporales. *Estadística Española*, Vol. 44, No. 150, 133 -159

[4] Alonso, A. M., Peña, D., y Romo, J. (2002). Forecasting time series with sieve Bootstrap. *Journal of Statistical Planning and Inference*, 100, 1-11.

[5] Alonso, A. M. y Romo, J. (2001). Forecast of the expected Nonpidemic Morbidity of Acure Diseases Using Resampling Methods. *Statistics and Econometrics Series*, 22, 1-34.

[6] Alonso, A. M., Peña, D., y Romo, J. (2000). Resampling time series by missing values techniques. *Working Paper* 00-42, Universidad Carlos III de Madrid, Madrid.

[7] Andrews W. K. y Buchinsky M. (1997). On the number of bootstrap repetitions for bootstrap standard errors, confidence intervals and tests. *Cowles Foundation Paper* 1141R.

[8] Arcones, M. y Giné, E. (1989). The bootstrap of the mean with arbitrary bootstrap sample size. *Annales de l'isnstitut Henri Poincaré*, 25,4. 457-481.

[9] Box, G. E. P. y Jenkins, G. M. y Reinsel, G. (1994). *Time Series Analysis Forecasting and Control*. Third Edition. Englewood Cliffs: Prentice Hall.

[10] Box, G. E. P. y Jenkins, G. M. (1976). *Time Series Analysis Forecasting and Control*. San Francisco: Holden-Day.

[11] Brockwell, P.J. y Davis, R.A. (2002). *Introduction to Time Series and Forecasting*, Springer Verlag, New York.

[12] Brockwell, P.J. y Davis, R.A. (1987). *Time Series Theory and Methods*, Springer Verlag, New York.

[13] Bühlmann, P. (2002). Bootstrap for time series. *Statistical Science*, 17, 52-72.

[14] Bühlmann, P. (1998). Sieve bootstrap for smoothing in non-stationary time series. *Annals Statistics*, 26, 48-83.

[15] Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, 3, 123-148.

[16] Cao, R. (1999). An overview of bootstrap methods for estimating and predicting time series. *Sociedad Española de Estadística e Investigación Operativa: Test*, 8, 95-116.

[17] Cao, R., M. Febrero-Bande, W. González-Manteiga, J.M. Prada-Sánchez, I. y García-Jurado (1997). Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes. *Comm. Statistic Simulation Computer*, 26, 961–978.

[18] Davidson, R. y MacKinnon, J. G. (1997). Bootstrap test: how many bootstraps? *Unpublished working paper*, Department of Economics, Queen's University, Kingston Ontario.

[19] Davison, A. C. y Hinkley, D. V. (1997). *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.

[20] Efron, B. y Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

[21] Efron, B. (1987).Better bootstrap confidence intervals, (with discussion) *Journal of the American Statistical Association*, 82, 171-200.

[22] Efron, B. y Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science*, 1, 54-77.

[23] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.

[24] Freedman, D. (1981). Bootstrapping regression models. *Annals of Statistics*, 9, 1218-1228.

[25] Grenander, U. (1981). *Abstract Inference*. New York: Wiley.

[26] Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, 14, 1453-1462.

[27] Hamilton, J.D (1994). *Time Series Analysis*. Princeton: Princeton University Press

[28] Hurvich, C. and Tsai, C. (1989). Regression and time series models selection in small samples, *Biometrika*, 76, 297-307.

[29] Jenkins, G. M. y Watts, D.G. (1968). *Spectral Analysis and Its Applications.* San Francisco: Holden-Day.

[30] Kreiss, J.P. y Franke, J. (1992). Bootstrapping stationary autoregressive moving-average models. *Journal of Time Series Analysis*, 13, 297-317.

[31] Kunsch, H.R. (1989). The Jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17, 1217-1241.

[32] Lahiri, S.N. (2003). *Resampling Methods for Dependent Data*, Series: Springer Series in Statistics, New York.

[33] Li, H. y Maddala, G.S. (1996). Bootstrapping time series models. *Econometric Reviews*, 15, 115-158.

[34] Morettin, P.A. y Toloi, C.M.C. (2004). *Análise de Séries Temporais.* São Paulo: Associação Brasileira de Estatística, no prelo.

[35] Morettin, P.A. (2002). *Econometria Financiera um Curso em Series Temporais Financeiras.* São Paulo: Instituto de Matemática e Ciencias Afins, IMCA, and Universidad Católica do Perú.

[36] Morettin, P.A. y Toloi, C.M.C. (1981). *Modelos para Previsão de Séries Temporais.* Rio de Janeiro: Instituto de matemática pura e aplicada.

[37] Politis, D.N. (2003). The Impact of Bootstrap Methods on Time Series Analysis. *Statistical Science*, Vol. 18, No. 2, 219-230.

[38] Politis, D.N. y Romano, J.P. y Wolf, M. (1999). *Subsampling*, Springer-Verlag, New York.

[39] Politis, D.N. y Romano, J.P. (1994). The Stationary Bootstrap. *Journal of American Statistic Association*, 89, 1303-1313.

[40] Ramírez, N. y Sastri, T. (1997). Transient Detection with an application to a Chemical Process. *Computers & Industrial Engineering,* Vol 32 No 4, pp 891-908.

[41] Rossow, W.B. y Schiffer, R.A. (1999). Advances in understanding clouds from ISCCP. *Bulletin of American Meteorological Society*, 80, 2261-2287..

[42] Shao, J. y Tu, D. (1995). *The Jackknife and Bootstrap*, Springer-Verlag, New York.

[43] Shibata, R. (1980). Asymptotically efficient selection of the order of the model estimating parameters of a lineal process, *Annals of Statistics*, 8, 147–164.

[44] Singh, K. (1981). On the Asymptotic accuracy of Efron's Bootstrap, *Annals of Statistics*, 9, 1187–1195.

[45] Thombs, L. A. y W. R. Schucany (1990). Bootstrap prediction intervals for autoregression. *Journal American Statistical Association*, 85, 486–492.

[46] Thombs, L. A. y W. R. Schucany (1990). Bootstrap prediction intervals for autoregression. *Journal American Statistical Association*, 85, 486–492.

[47] Uriel, J. E. (1985). *Análisis de Series Temporales modelos ARIMA*, Paraninfo colección Ábaco, Madrid.

APENDICE A

Suposiciones y Resultados Principales del Sieve Bootstrap.

Consideremos con más detalle los modelos (4.2) y (4.3) para dar propiedades acerca del proceso estacionario $\{X_t\}_{t\in\mathbb{Z}}$, del cual se extrajo una muestra $X_1,\ldots,X_n$; usualmente para describir las propiedades de este proceso es mejor representarlo mediante un modelo $MA(\infty)$, así para los modelos (4.2) y (4.3), tenemos

$$\Phi(z)=\sum_{j=o}^{\infty}\phi_i z^j,\quad \phi_0=1,\ z\in\mathbb{C}$$

$$\Psi(z)=\sum_{j=o}^{\infty}\psi_i z^j,\quad \psi_0=1,\ z\in\mathbb{C}$$

Los respectivos polinomios, y podemos escribirlos en sus formas compactas como

$$\Phi(B)(X-\mu_X)=\varepsilon$$
$$X-\mu_X=\Psi(B)\varepsilon$$

Respectivamente, donde B es el operador de retardos, así formalmente tenemos que $\Psi(z)=1/\Phi(z)$, sea $\mathbf{F}_t=\sigma(\{\varepsilon_s;\ s\le t\})$ σ-campo generado por $\{\varepsilon_s\}_{s=-\infty}^{t}$, entonces

(1) $X_t-\mu_X=\sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$, $\psi_0=1$, $(t\in\mathbb{Z})$, con $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ estacionario, ergódico y

$E[\varepsilon_t\mid\mathbf{F}_{t-1}]\equiv 0$, $E[\varepsilon_t^2\mid\mathbf{F}_{t-1}]\equiv\sigma^2<\infty$, $E[\varepsilon_t]^s<\infty$, para $s\ge 4$.

(2) El polinomio $\Psi(z)$ es acotado, esto es que para $|z|\le 1$, $\sum_{j=0}^{\infty} j^r|\psi_j|<\infty$, para algún

$r\in\mathbb{N}$.

Esta suposición incluye a modelos con coeficientes que tiene un decaimiento polinomial $\{\psi_j\}_{j=0}^{\infty}$ o equivalentemente $\{\phi_j\}_{j=0}^{\infty}$. Los modelos *ARMA(p,q)* que tienen esta suposición tiene un decaimiento exponencial de $\{\psi_j\}_{j=0}^{\infty}$. Finalmente esta suposición implica que el polinomio $\Phi(z)$ es acotado, esto es que $|z| \leq 1, \sum_{j=0}^{\infty} j^r |\phi_j| < \infty$.

Puesto que el esquema del *sieve bootstrap* extrae independientemente de los residuales, luego éste usualmente es incapaz de reproducir la estructura de probabilidad de un estadístico basado en el modelo (1) con variables $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ no independientes. Una excepción al respecto es la media aritmética por ser un estadístico lineal. Algunas veces se consolida (1) como

(1') $$X_t - \mu_X = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \psi_0 = 1, (t \in \mathbb{Z}), \quad \text{con} \quad \{\varepsilon_t\}_{t\in\mathbb{Z}} \quad \text{iid}$$

$$E[\varepsilon_t] = 0, \quad E[\varepsilon_t]^s < \infty, \text{ para } s \geq 4.$$

Esta suposición es más restrictiva ya que considera los errores independientes e idénticamente distribuidos. Otro punto importante es el caso de la aproximación autorregresiva que usa el siguiente criterio

(I) $p = p(n) \to \infty, \quad p(n) = o(n), \quad n \to \infty$ y $\hat{\phi}_p = \left(\hat{\phi}_{1,n}, \hat{\phi}_{2,n}, \ldots, \hat{\phi}_{p,n}\right)^T$ satisface las ecuaciones de *Yule Walker*

$$\hat{\Gamma}_p \hat{\phi}_p = -\hat{\gamma}_p$$

Donde $\hat{\Gamma}_p = \left[\hat{\gamma}(i-j)\right]_{i,j=1}^{p}$, $\hat{\gamma}_p = (\hat{\gamma}(1),\ldots,\hat{\gamma}(p))^T$, con

$$\hat{\gamma}(h) := (1/n)\sum_{j=1}^{n-h}(x_{j+h}-\bar{x})(x_j-\bar{x}), \quad 0 \le h < n.$$

Los resultados de consistencia del *sieve bootstrap*, en el simple caso de que el estimador *sieve bootstrap* sea la varianza de la media, aun considerando un proceso como en (1) con innovaciones no independientes. Son mostrados y comparados con el *bootstrap* por bloques son dados en Bühlmann, 1997. Así como para clases de estimadores no lineales, considerando (1').

Propiedades del Sieve Bootstrap

- **Aproximación Autorregresiva**: usando el procedimiento de estimación como en (I)**,** $\hat{\Phi}_n(z) = \sum_{j=0}^{p(n)} \hat{\phi}_{j,n} z^j$, $\hat{\phi}_{0,n} = 1$, $(z \in \mathbb{C}, |z| \le 1)$, se sabe que $\hat{\Phi}_n(z)$ es invertible para $|z| \le 1$, esto es que $1/\hat{\Phi}_n(z) = \hat{\Psi}_n(z) = \sum_{j=0}^{\infty} \hat{\psi}_{j,n} z^j$, $(|z| \le 1)$, así la representación *sieve bootstrap* para (4.2) es $X_t^* - \bar{X} = \sum_{j=0}^{\infty} \hat{\psi}_{j,n} \varepsilon_{t-j}^*$, $t \in \mathbb{Z}$.. Asumiendo (1) con $s = 4$, (2) con $r = 1$ y (I) con $p(n) = o\left((n/\log(n))^{1/4}\right)$, entonces:

$\sup_{0 \le j < \infty} |\hat{\psi}_{j,n} - \psi_j| = o(1)$, $n \to \infty$**,** casi seguramente.

- **Sieve Bootstrap Muestral**: inicialmente se presenta algunos resultados acerca del remuestreo en las innovaciones ε_t^* $i.i.d.$ $\hat{F}_{\varepsilon,n}$, por la definición de $\hat{F}_{\varepsilon,n}$ se tiene $E_*(\varepsilon_t^*) = 0$; y para momentos altos, se tiene. Asumiendo (1) con

$s = \max\{2w, 4\}$, $w \in \mathbb{N}$, (2) con $r = 0$ y (I) con $p(n) = o\left((n/\log(n))^{1/2}\right)$, entonces: $E_*\left[(\varepsilon_t^*)^{2w}\right] = E\left[(\varepsilon_t)^{2w}\right] + o_P(1)$, donde o_P denota el orden de convergencia en probabilidad. Por otro lado asumiendo (1) con $s = 4$, (2) con $r = 1$ y (I) con $p(n) = o\left((n/\log(n))^{1/4}\right)$, entonces: $\varepsilon_t^* \xrightarrow{d^*} \varepsilon_t$ en probabilidad, donde d^* denota la función de distribución. Consecuentemente asumiendo (1) con $s = 4$, (2) con $r = 0$ y (I) con $p(n) = o\left((n/\log(n))^{1/2}\right)$, análogamente entonces:

$X_t^* \xrightarrow{d^*} X_t$ en probabilidad

Así las propiedades del *sieve bootstrap* hacen consistentes las estimaciones bajo supuestos establecidos

La validez asintótica

Para el Intervalo de confianza (4.11) el comportamiento límite de la distribución $F^*_{X_{T+h}}$, y para esto es suficiente establecer convergencia en distribución condicional de la versión *bootstrap* X^*_{T+h} a X_{T+h}. Este procedimiento *sieve bootstrap* tiene dos principales partes:

- Obtener las estimaciones $\hat{\phi}_p^*$ en orden para tener información acerca de la distribución de $\hat{\phi}_p$.
- Calcular los valores futuros X^*_{T+h}.

Para conseguir estas condiciones, se consideran previas suposiciones acerca del proceso estacionario $\{X_t\}_{t\in\mathbb{Z}}$ las cuales son:

(3) $X_t - \mu_X = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \psi_0 = 1, \qquad (t \in \mathbb{Z}), \qquad$ con $\qquad \{\varepsilon_t\}_{t\in\mathbb{Z}} \qquad$ i.i.d. y

$E[\varepsilon_t \mid \mathbf{F}_{t-1}] \equiv 0, \;\; E[\varepsilon_t^2 \mid \mathbf{F}_{t-1}] \equiv \sigma^2 < \infty, \;\; E[\varepsilon_t]^s < \infty$, para $s \geq 4$ y $\mathbf{F}_{t-1}$ σ-campo generado por $\{\varepsilon_s\}_{s=-\infty}^{t-1}$.

Adicionalmente las suposiciones (2) y (I), entonces se tienen los siguientes resultados

Proposición.- Supongamos las suposiciones (3), con $s = 4$, (2), con $r > 2$, y (I) con $p = o\left((n/\log n)^{1/(2r+2)}\right)$, entonces:

$\max_{0\leq j<p(n)} |\hat{\phi}_j^* - \hat{\phi}_j| \xrightarrow{P^*} 0$, en probabilidad

Teorema.- Supongamos las suposiciones (3), con $s = 4$, (2), con $r = 1$, y (I) con $p = o\left((n/\log n)^{1/4}\right)$, entonces:

$X_{T+h}^* \xrightarrow{d} X_{T+h}$, en probabilidad

Estos resultados sirven para validar la convergencia de la versión *bootstrap* X_{T+h}^* a X_{T+h}, para detalles ver [Alonso, 2003].

La Distancia de Mallows

Mallows en 1972 propone una medida de diferencia entre dos distribuciones de probabilidad *P* y *Q*, ambas definidas sobre $\mathbb{R}^n$, la cual esta dada por:

$Mallow_p(P,Q) = \min_{\mu}(E_\mu \| x - y \|_p^p)^{1/p}$,

Sujeto a $\int_y d\mu(x,y) = P(x), \int_x d\mu(x,y) = Q(x)$

Donde $\|\cdot\|_p$ denota la norma L_p y $1 \leq p < \infty$

Programa SAMPLE: Ejemplo como utilizar las subrutinas.

- Subrutina D2OPEN: Abre un archivo D2 e inicializa.
- Subrutina D2READ: Desempaca la data D2 para una banda de latitud en cantidades de valores enteros.
- Subrutina D2REC: Usado por D2READ para desempacar un registro lógicamente.
- Subrutina D2PHYS: Convierte cantidades enteras en una banda de latitud a valores físicos.
- Subrutina RDANC: Lee los archivos de datos auxiliares.
- Subrutina PRINTI: Imprime parámetros de cantidad de valores para un recuadro grid.
- Subrutina PRINTR: Imprime valores físicos para un recuadro grid.
- Subrutina CENTER: Calcula el centro latitud/longitud de una celda grid.
- Subrutina CLDHGT: Calcula la altura en metros de la parte superior de la nube.
- Subrutina EQ2SQ: Convierte un mapa de igual área en un mapa de igual ángulo.
- Subrutina BLOCK DATA: Conversión a tablas de la información de un grid.

APENDICE B

Las Rutinas Creadas en MATLAB para la Experimentación.

Se muestran a continuación todas las rutinas en MATLAB, que permiten estimar los intervalos de predicción bajo el método del *Sieve Bootstrap* de una o más series de tiempo, seleccionando un modelos autorregresivos, realizando remuestreo en los residuales, evaluando los residuales en los modelos seleccionados y construyendo intervalos de predicción para diferentes pasos adelante, así como para diferentes criterios de elección de modelos y diferentes tamaños de las remuestras en los residuales.

- **fsieve.**

```
% Routine to generate the empirical forecast distribution
% for an SARIMA(p,d,0)(ps,ds,0)
% Input:
% ------
% XSeries : time series vector.
% T : time series size.
% d, ds, season : model's structure.
% kfc : parameter used by routine g13bef.
% nreplicas : number of  forecast used to estimate efd.
% maxlag : maximum lag to forecast.
% criteria : information criteria for selecting AR model.

function [R1, R2, R3] = fsieve(XSeries, T, d1, ds1, season, kfc, ...
nreplicas, maxlag, criteria)
XMean = mean(XSeries);
YSeries = XSeries - XMean;
if (d1 == 0)
   kfc = 0;
end
pmax = round(min(T/10, 10*log10(T)));
if (season > 1)
```

```
  Pmax = 3;
else
  Pmax = 0;
end
[phat, pshat, Sres, Spara] = armselect(YSeries, T, ...
  pmax, d1, Pmax, ds1, season, kfc, criteria);
if (phat > 0)
  NewPhi = Spara(1:phat,1)';
else
  NewPhi = [];
end
if (pshat > 0)
  NewPHI = Spara(phat+1:phat+pshat,1)';
else
  NewPHI = [];
end
Cmodel = Spara(phat+pshat+1, 1)*kfc;
EResiduals = [zeros(d1+ds1*season,nreplicas); ...
    nncopy(Sres(1:T-d1-ds1*season),1,nreplicas); ...
    zeros(maxlag,nreplicas)]';
WSeries = [nncopy(YSeries,nreplicas,1) zeros(nreplicas,maxlag)];
TruePhi = zeros(1,pshat*season+1);
TruePhi(1) = 1;
for i = 1:pshat
  TruePhi(i*season+1) = -NewPHI(i);
end
TruePhi = conv(TruePhi, [1,-NewPhi]);
if d1 == 1   % We only consider d or ds = 0/1.
  TruePhi = conv(TruePhi, [1 -1]);
end
if d1 == 2   % We only consider d or ds = 0/1.
  TruePhi = conv(TruePhi, [1 -1]);
  TruePhi = conv(TruePhi, [1 -1]);
end
if ds1 == 1
  TrueDS = zeros(1,ds1*season+1);
  TrueDS(1) = 1;
  TrueDS(ds1*season+1) = -1;
  TruePhi = conv(TruePhi, TrueDS);
end
EResiduals(1:nreplicas, T+1:T+maxlag) = ...
  rngnag(nreplicas, maxlag, 'empirical', Sres(1:T-d1-ds1*season));
WSeries(1:nreplicas, T+1:T+maxlag) = ...
```

```
  EResiduals(1:nreplicas, T+1:T+maxlag);
for t = T+1:T+maxlag
  WSeries(1:nreplicas, t) = Cmodel + WSeries(1:nreplicas, t);
  for ip = 2:(phat+d1+(pshat+ds1)*season+1)
    WSeries(1:nreplicas, t) = WSeries(1:nreplicas, t) ...
      - TruePhi(ip)*WSeries(1:nreplicas, t-ip+1);
  end
end
R1 = XMean + WSeries(1:nreplicas,T+1:T+maxlag);
R2 = Sres(1:T-d1-ds1*season);
R3 = Spara(1:phat+pshat+1);

% Output:
% -------
% R1 : bootstrap forecast distribution (nreplicas x maxlag matrix).
% R2 : estimated residuals (T-d-ds*season x 1 vector).
% R3 : estimated parameters (p+ps x 1 vector).
```

- **armselect.**

```
% Routine to select the approximating ARIMA(phat, d1, 0)(pshat, ds1)
% model to the given XSeries.
% Input:
% ------
% XSeries : time series vector.
% T : time series size.
% pmax, Pmax : maximum regular and seasonal autoregressive orders.
% d1, ds1, season : fixed model parameters.
% kfc : parameter used by routine g13bef.
% criteria : information criteria (aic,aicc, bic)

function [R1, R2, R3, R4] = armselect(XSeries, T, ...
  pmax, d1, Pmax, ds1, season, kfc, criteria)
icvalue = ones(pmax+1, Pmax+1);
for i = 0:pmax
  for j = 0:Pmax
    if (i+j == 0)
      [res, Tres, ifail] = g13aaf(XSeries, d1, ds1, season);
      icvalue(1, 1) = var(res);
      para = mean(res);
    else
      kef = 1; % for 1= least-square, 2= exact likelihood,3= marginal likelihood
```

```
        kzef = 0;
        kzsp = 0;
        zsp = [0.01; 10.0; 1000.0; 0.0001];
        [para,xxy,itc,sd,cm,s,d,ndf,res,sttf,nsttf,zsp,ifail] = ...
          g13bef([i d1 0 j ds1 0 season], [0;0;0;1], ...
          zeros(1, i+j+1), XSeries', kzsp, zsp, kfc);
        icvalue(i+1, j+1) = s/T;
      end
    end
  end
  switch lower(criteria)
          case {'aic'}
    for i = 0:pmax
      for j = 0:Pmax
        icvalue(i+1, j+1) = log(icvalue(i+1, j+1)) + 2*(i+j)/T;
      end
    end
          case {'aicc'}
    for i = 0:pmax
      for j = 0:Pmax
        icvalue(i+1, j+1) = log(icvalue(i+1, j+1)) + (T+i+j)/(T-i-j-2);
      end
    end
    case {'bic'}
      for i = 0:pmax
        for j = 0:Pmax
          icvalue(i+1, j+1) = log(icvalue(i+1, j+1)) + log(T)*(i+j)/T;
        end
      end
    end
  if (Pmax > 0)
    [icmin, pshat] = min(min(icvalue));
    [icmin, phat] = min(min(icvalue'));
    phat = phat-1;
    pshat = pshat-1;
  else
    [icmin, phat] = min(icvalue);
    phat = phat-1;
    pshat = 0;
  end
  kef = 1; % for using least-square
  kzef = 0;
  kzsp = 0;
```

```
zsp = [0.01; 10.0; 1000.0; 0.0001];
if (phat + pshat > 0)
  [para,xxy,itc,sd,cm,s,d,ndf,res,sttf,nsttf,zsp,ifail] = ...
    g13bef([phat d1 0 pshat ds1 0 season], [0;0;0;1], ...
    zeros(phat+pshat+1, 1), XSeries', kzsp, zsp, kfc);
else
  [res, Tres, ifail] = g13aaf(XSeries, d1, ds1, season);
  para = mean(res);
end
R1 = phat;
R2 = pshat;
R3 = res;
R4 = para(1:phat+pshat+1);

% Output:
% ---------
% R1 : selected p = phat.
% R2 : selected P = pshat.
% R3 : estimated residuals (T-d-ds*season x 1 vector).
% R4 : estimated parameters (phat+pshat x 1 vector).
```

- **rngnag**

```
% Routine to generate pseudo-random numbers (with zero mean)
% Input:
% ------
% nr, nc : matrix's size.
% edist : error's distribution.
% edp : error's distribution parameters.

function Results = rngnag(nr, nc, edist, edp)
Xrandom = zeros(nr, nc);
switch lower(edist)
  case {'normal'}
    for i = 1:nc
      Xrandom(1:nr, i) = g05fdf(edp(1), edp(2), nr);
    end
  case {'gamma'}
    for i = 1:nc
      Xrandom(1:nr, i) = g05fff(edp(1), edp(2), nr) - edp(1)*edp(2);
    end
  case {'uniform'}
```

```
    for i = 1:nc
      Xrandom(1:nr, i) = g05faf(edp(1), edp(2), nr) - (edp(1)+edp(2))/2;
    end
  case ('t')
    Xrandom = trnd(edp(1), nr, nc);
  case {'empirical'}
    Yrandom = edp - mean(edp);
    Xorders = unidrnd(max(size(edp)), nr, nc);
    for i = 1:nc
      Xrandom(1:nr, i) = Yrandom(Xorders(1:nr, i));
    end
end
Results = Xrandom;

% Output:
% -------
% Results is a nr x nc vector.
```

- **ts**

```
% TS command executes TRAMO-SEATS for a set of time series
% INPUT:
%  x:    An n*m matrix containing the set of time series to be processed.
%  nom:  A cell array of strings containing the names of the series,
%  year: Date of the beginning of the series.
%  per:  Initial period for the beginning of the series.
%  s:    Frequency of the data.

function [res]=ts(x,nom,year,per,s,opc,varargin)
pathTramo = 'c:\tramo'; % The TRAMO directory
pathSeats = 'c:\seats'; % The SEATS directory
executable = 'c:\tramo\ts.exe';
cadts = 'res = tsml1(x,nom,year,per,s,pathTramo,pathSeats,executable,opc';
if nargin>6
   for i=1:nargin-6
      cadts = [cadts ',' 'varargin{' num2str(i) '}'];
   end
end
cadts = [cadts ');'];
eval(cadts)

% OUTPUT:
```

% A m structure arrays vector. Each one of the m structure arrays contains fixed fields:
% · name of the series: name
% · the original series: xorig
% · a series with the dates: date
% · a structure with the model estimated by TRAMO: model
% · a structure with the main descriptive statistics about the residuals: stat
% · a structure containing the ARMA coefficients estimated: arma
% · a structure with the (total) deterministic effects: dete
% · a structure with the calendar effects: cal
% · a structure with the outlier effects: out
% · a structure with the regression variables: vreg
% - Set of variable fields: the series generated y TRAMO-SEATS for each one of
% the m series analyzed.

- **fsieve_nubes.**

% Routine to construct confidence intervals for clouds data using to TRAMO-SEATS for
% select the model, then using fsieve.
% ----------
% Clouds data on Caribe:
% ----------
% SCA Stratocumulus Amount (%)
% SCTP Stratocumulus Top Pressure (milibars)
% SCTT Stratocumulus Top Temperature (kelvin)
% SCOT Stratocumulus Optical Thickness (tau)
% SCWP Stratocumulus Water Path (galon/metro^2)
% SA Stratus Amount (%)
% STP Stratus Top Pressure (milibars)
% STT Stratus Top Temperature (kelvin)
% SOT Stratus Optical Thickness (tau)
% SWP Stratus Water Path (galon/metro^2)
% NSA Nimbustratus Amount (%)
% NSTP Nimbustratus Top Pressure (milibars)
% NSTT Nimbustratus Top Temperature (kelvin)
% NSOT Nimbustratus Optical Thickness (tau)
% NSWP Nimbustratus Water Path (galon/metro^2)
% DCA Deep convection Amount (%)
% DCTP Deep Convection Top Pressure (milibars)
% DCTT Deep Convection Top Temperature (kelvin)
% DCOT Deep Convection Optical Thickness (tau)
% DCWP Deep Convection Water Path (galon/metro^2)

```
clear all, close all,clc,
XData =load ‘clouds.txt’;
SeriesNames = ['SCA '; 'SCTP'; 'SCTT'; 'SCOT'; 'SCWP'; 'SA  '; 'STP '; 'STT '; 'SOT '; 'SWP
'; 'NSA '; 'NSTP'; ...
    'NSTT'; 'NSOT'; 'NSWP'; 'DCA '; 'DCTP'; 'DCTT'; 'DCOT'; 'DCWP'];
[T ng] = size(XData);

% Options for TRAMO-SEATS for select the orders from models.
opts = 'lam=-1,itrad=-1,ieast=-1,idur=9,inic=3,idif=3,iatip=1,aio=2,ireg=0';

% Selecting the model by TRAMO-SEATS.
d1 = zeros(1, ng); % Number of regular differences.
ds1 = zeros(1, ng); % Number of seasonal differences.
kfc = zeros(1, ng); % Null or non null constant.
lam = zeros(1, ng); % Log transformation.

% Running TRAMO-SEATS by ts function.
for i = 1:ng
  display(['Series No. ', num2str(i)])
  sal1 = ts(XData(:,i),{SeriesNames(i,:)},1983,7,12,opts);
  d1(1, i) = sal1.model.d;
  ds1(1, i) = sal1.model.bd;
  kfc(1, i) = sal1.model.mean;
  lam(1, i) = sal1.model.lam;
end

% Parameters for fsieve.
B = 1000;  % Number of Resampling
maxlag = 12;   % Maximum prediction lag. Ahead
criteria = 'aic'    %criteria of select models.;
season = 12;

% Predictions for cloud data using sieve bootstrap
SBefd = zeros(B, maxlag, ng);
for i = 1
  if (lam(i) == 1)
    [SBefdB SBresB SBparaB] = fsieve(XData(1:T,i)', T, d1(i), ds1(i), season, kfc(i), ...
      B, maxlag, criteria);
    SBefd(:, :, i) = SBefdB;
  else
    [SBefdB SBresB SBparaB] = fsieve(log(XData(1:T,i))', T, d1(i), ds1(i), season, kfc(i), ...
      B, maxlag, criteria);
    SBefd(:, :, i) = exp(SBefdB);
```

```
  end
end


% plotting the predictions and confidence intervals for all cloud data
for i = 1:4
  for j = 1:5
    figure(1)
    subplot(4,5,(i-1)*5+j)
    plot(1:T, XData(:,(i-1)*5+j), 'b-')
    title(SeriesNames((i-1)*5+j,1:4));
    hold on
    plot(T+1:T+12, mean(squeeze(SBefd(:,:,(i-1)*5+j))), 'r-')
    SortedSBefd = sort(squeeze(SBefd(:,:,(i-1)*5+j)));
    plot(T+1:T+12, SortedSBefd(.975*B, :), 'g-')     % level of significance 5%
    plot(T+1:T+12, SortedSBefd(.025*B, :), 'g-')     % level of significance 5%
  end
end


% plotting the predictions and confidence intervals for each one features from cloud data
for k =1:ng
 figure(k)
 plot(1:T,XData(:,k))
 title(SeriesNames((i-1)*5+j,1:4));
 hold on
 plot(T+1:T+12, mean(squeeze(SBefd(:,:,k))), 'r-')
  SortedSBefd = sort(squeeze(SBefd(:,:,k)));
  plot(T+1:T+12, SortedSBefd(.975*B, :), 'g-')       % level of significance 5%
  plot(T+1:T+12, SortedSBefd(.025*B, :), 'g-')       % level of significance 5%
 grid on
 figure(k+1)
 plot(T+1:T+12, mean(squeeze(SBefd(:,:,k))), 'r-')
 hold on
 plot(T+1:T+12, SortedSBefd(.975*B, :), 'g-')       % level of significance 5%
 plot(T+1:T+12, SortedSBefd(.025*B, :), 'g-')       % level of significance 5%
 grid on
end
```

Las rutinas modificadas de FORTRAN, que se muestran a continuación, permiten seleccionar los datos D2 del *ISCCP* según el área de trabajo durante el periodo de julio/1987 a diciembre/2004.

- **D2readfinal**

```
C*              D 2 R E A D   M O D U L E
C* before running, link the input files to the fortran units:
C*    ln -s <d2ancilfile> fort.9
C*    ln -s <d2datafile> fort.10
C*    PROGRAM SAMPLE    :  EXAMPLE OF HOW TO USE THESE SUBROUTINES
C*    SUBROUTINE D2OPEN :  OPEN A D2 FILE AND INITIALIZE
C*    SUBROUTINE D2READ :  UNPACK D2 DATA FOR ONE LATITUDE BAND
C*    SUBROUTINE D2REC  :  USED BY D2READ TO UNPACK A LOGICAL RECORD
C*    SUBROUTINE D2PHYS :  CONVERT DATA IN LAT BAND TO PHYSICAL VALUES
C*    SUBROUTINE MIDPRS :  CALCULATE MID-LAYER PRESSURES FOR GRID BOX
C*    SUBROUTINE RDANC  :  READ ANCILLARY DATA FILE
C*    SUBROUTINE PRINTI :  PRINT COUNT VALUES FOR ONE GRID BOX
C*    SUBROUTINE PRINTR :  PRINT PHYSICAL VALUES FOR ONE GRID BOX
C*    SUBROUTINE CENTER :  CALCULATE CENTER LON/LAT OF GRID BOX
C*    SUBROUTINE CLDHGT :  CALCULATE CLOUD TOP HEIGHT IN METERS
C*    SUBROUTINE EQ2SQ  :  CONVERT EQUAL AREA MAP TO SQUARE MAP
C*    BLOCK DATA        :  CONVERSION TABLES AND EQUAL-AREA GRID INFO
      PROGRAM SAMPLE
      PARAMETER    ( MAXVAR = 130 )
      PARAMETER    ( MAXLON = 144 )
      PARAMETER    ( MAXLAT = 72 )
      PARAMETER    ( MAXBOX = 6596 )
      PARAMETER    ( IUNDEF = 255 )
      PARAMETER    ( RUNDEF = -1000.0 )
      COMMON /D2HEAD/ LUND2,IREC,IFILE,IYEAR,MONTH,IDAY,IUTC
     $    ,LATBEG,LATEND,LONBEG,LONEND,IBXBEG,IBXEND
      COMMON /D2DATA/ LAT,NLON,IVAR(MAXVAR,MAXLON),RVAR(MAXVAR,MAXLON)
      COMMON /D2GRID/ NCELLS(MAXLAT),ICELLS(MAXLAT)
      PARAMETER    ( MAXCNT = 255 )
      COMMON/CNTTAB/TMPTAB(0:MAXCNT),TMPVAR(0:MAXCNT),PRETAB(0:MAXCNT),
     1          RFLTAB(0:MAXCNT),TAUTAB(0:MAXCNT),PRWTAB(0:MAXCNT),
     2          OZNTAB(0:MAXCNT)
      REAL*4 EQMAP(MAXBOX)
      REAL*4 SQMAP(MAXLON,MAXLAT)
      CHARACTER*100 STARS/'************************************************
     $*****************************************************'/
      LUNANC = 9
      CALL RDANC(LUNANC,IRC)
      IF ( IRC .NE. 0 ) GOTO 900
      DO I=1,MAXBOX
        EQMAP(I) = -1000.0
      END DO
      LUND2 = 10
      CALL D2OPEN(IRC)
      IF ( IRC .NE. 0 ) GOTO 910
```

```
      IBOX = 0
      IFULL = 0
      DO 500 LAT=1,MAXLAT
        CALL D2READ(IRC)
        IF ( IRC .LT. 0 ) THEN
          GOTO 800
        ELSE IF ( IRC .GT. 0 ) THEN
          GOTO 920
        END IF
         CALL D2PHYS
        DO 400 LON=1,NLON
        IBOX = IBOX + 1
        IF ( IVAR(5,LON) .EQ. 255 ) GOTO 400
        IFULL = IFULL + 1
C* LOCALIZATION: CARIBBEAN ISLANDS
      IF (LAT.LT.47.AND.LAT.GT.42.AND.LON.LT.115.AND.LON.GT.100)THEN
          CALL PRINTR(LON)
        END IF
        EQMAP(IBOX) = RVAR(8,LON)
  400   CONTINUE
  500 CONTINUE
  800 CONTINUE
      CALL EQ2SQ(1,EQMAP,SQMAP)
      LUNOUT = 90
      OPEN(LUNOUT,ACCESS='DIRECT',RECL=576,FORM='UNFORMATTED')
      DO 850 J=1,MAXLAT
        WRITE(LUNOUT,REC=J) (SQMAP(I,J),I=1,MAXLON)
  850 CONTINUE
      STOP 0
  900 CONTINUE
      PRINT *,'ERROR:  RDANC  RC=',IRC
      STOP 999
  910 CONTINUE
      PRINT *,'ERROR:  D2OPEN RC=',IRC
      STOP 999
  920 CONTINUE
      PRINT *,'ERROR:  D2READ RC=',IRC
      STOP 999
      END
      SUBROUTINE D2OPEN(IRC)
      COMMON /D2HEAD/ LUND2,IREC,IFILE,IYEAR,MONTH,IDAY,IUTC
     $    ,LATBEG,LATEND,LONBEG,LONEND,IBXBEG,IBXEND
      IBXBEG = 0
      IBXEND = 0
      IREC = 0
      OPEN(LUND2,ACCESS='DIRECT',RECL=13000,
     $    FORM='UNFORMATTED',IOSTAT=IRC)
      RETURN
      END
      SUBROUTINE D2READ(IRC)
      PARAMETER    ( MAXVAR = 130 )
      PARAMETER    ( NUMBOX = 100 )
      PARAMETER    ( MAXLAT =  72 )
      PARAMETER    ( MAXLON = 144 )
      PARAMETER    ( IUNDEF = 255 )
      PARAMETER    ( RUNDEF = -1000.0 )
      COMMON /D2BUFS/ CHRBUF(MAXVAR,NUMBOX)
      CHARACTER*1     CHRBUF
```

```
      COMMON /D2DATA/ LAT,NLON,IVAR(MAXVAR,MAXLON),RVAR(MAXVAR,MAXLON)
      COMMON /D2HEAD/ LUND2,IREC,IFILE,IYEAR,MONTH,IDAY,IUTC
     $     ,LATBEG,LATEND,LONBEG,LONEND,IBXBEG,IBXEND
      COMMON /D2GRID/ NCELLS(MAXLAT),ICELLS(MAXLAT)
      SAVE IDECOD
      DO 100 LON=1,MAXLON
      DO 100 I=1,MAXVAR
        IVAR(I,LON) = IUNDEF
        RVAR(I,LON) = RUNDEF
  100 CONTINUE
      NLON = ICELLS(LAT)
      NPREV = NCELLS(LAT)
      DO 500 LON=1,NLON
        NBOX = NPREV + LON
  200   CONTINUE
        IF ( NBOX .GE. IBXBEG ) THEN
          IF ( NBOX .LE. IBXEND ) THEN
            IF ( ICHAR(CHRBUF(1,IDECOD+1)) .GT. LAT ) GOTO 510
            IDECOD = IDECOD + 1
            ILON = ICHAR(CHRBUF(2,IDECOD))
            DO 300 I=1,MAXVAR
              IVAR(I,ILON) = ICHAR(CHRBUF(I,IDECOD))
  300         CONTINUE
          ELSE
            CALL D2REC(IRC)
            IDECOD = 1
            IF ( IRC .EQ. 0 ) THEN
              GOTO 200
            ELSE
              GOTO 900
            END IF
          END IF
        END IF
  500 CONTINUE
  510 CONTINUE
  900 CONTINUE
      RETURN
      END
      SUBROUTINE D2REC(IRC)
      PARAMETER    ( MAXVAR = 130 )
      PARAMETER    ( NUMBOX = 100 )
      PARAMETER    ( MAXLAT =  72 )
      COMMON /D2BUFS/ CHRBUF(MAXVAR,NUMBOX)
      CHARACTER*1     CHRBUF
      COMMON /D2HEAD/ LUND2,IREC,IFILE,IYEAR,MONTH,IDAY,IUTC
     $     ,LATBEG,LATEND,LONBEG,LONEND,IBXBEG,IBXEND
      COMMON /D2GRID/ NCELLS(MAXLAT),ICELLS(MAXLAT)
      IREC = IREC + 1
      READ(LUND2,REC=IREC,IOSTAT=IRC) CHRBUF
      IF ( IRC .EQ. 0 ) THEN
        JREC   = ICHAR(CHRBUF(1,1))
        IFILE  = ICHAR(CHRBUF(2,1))
        IYEAR  = ICHAR(CHRBUF(3,1))
        MONTH  = ICHAR(CHRBUF(4,1))
        IDAY   = ICHAR(CHRBUF(5,1))
        IUTC   = ICHAR(CHRBUF(6,1))
        LATBEG = ICHAR(CHRBUF(7,1))
        LONBEG = ICHAR(CHRBUF(8,1))
```

```
      LATEND = ICHAR(CHRBUF(9,1))
      LONEND = ICHAR(CHRBUF(10,1))
      IBXBEG = NCELLS(LATBEG) + LONBEG
      IBXEND = NCELLS(LATEND) + LONEND
      IF ( IREC .EQ. 1 )  PRINT 90,IYEAR,MONTH,IDAY,IUTC
90    FORMAT()
    END IF
    RETURN
    END
    SUBROUTINE EQ2SQ(ISHIFT,EQMAP,SQMAP)
    PARAMETER     ( MAXLON = 144 )
    PARAMETER     ( MAXLAT = 72 )
    PARAMETER     ( MAXBOX = 6596 )
    REAL EQMAP(MAXBOX)
    REAL SQMAP(MAXLON,MAXLAT)
    COMMON /D2GRID/ NCELLS(MAXLAT),ICELLS(MAXLAT)
    COMMON /SQUARE/ LONLIM(2,MAXBOX)
    IBOX = 0
    DO 200 LAT=1,MAXLAT
    DO 200 LON=1,ICELLS(LAT)
      IBOX = IBOX + 1
      LONSQ1 = LONLIM(1,IBOX)
      LONSQ2 = LONLIM(2,IBOX)
      DO 100 ILON=LONSQ1,LONSQ2
        LONSQ = ILON
        IF ( ISHIFT .EQ. 1 ) THEN
          LONSQ = LONSQ + MAXLON/2
          IF ( LONSQ .GT. MAXLON ) LONSQ = LONSQ - MAXLON
        END IF
        SQMAP(LONSQ,LAT) = EQMAP(IBOX)
100   CONTINUE
200 CONTINUE
    RETURN
    END
    SUBROUTINE D2PHYS
    PARAMETER     ( PATHW = 6.292 )
    PARAMETER     ( PATHI = 10.5 )
    PARAMETER     ( MAXVAR = 130 )
    PARAMETER     ( NUMBOX = 100 )
    PARAMETER     ( MAXLON = 144 )
    PARAMETER     ( IUNDEF = 255 )
    PARAMETER     ( RUNDEF = -1000.0 )
    COMMON /D2DATA/ LAT,NLON,IVAR(MAXVAR,MAXLON),RVAR(MAXVAR,MAXLON)
    PARAMETER     ( MAXCNT = 255 )
    COMMON/CNTTAB/TMPTAB(0:MAXCNT),TMPVAR(0:MAXCNT),PRETAB(0:MAXCNT),
   1            RFLTAB(0:MAXCNT),TAUTAB(0:MAXCNT),PRWTAB(0:MAXCNT),
   2            OZNTAB(0:MAXCNT)
    DO 500 LON=1,NLON
      DO 100 I=1,7
        IF ( IVAR(I,LON) .EQ. IUNDEF ) THEN
          RVAR(I,LON) = RUNDEF
        ELSE
          RVAR(I,LON) = FLOAT(IVAR(I,LON))
        ENDIF
100   CONTINUE
      DO 101 I=8,9
        IF ( IVAR(I,LON) .EQ. IUNDEF ) THEN
          RVAR(I,LON) = RUNDEF
```

```
        ELSE
          RVAR(I,LON) = FLOAT(IVAR(I,LON)) * 0.5
        ENDIF
101     CONTINUE
      DO 102 I=10,19
        IF ( IVAR(I,LON) .EQ. IUNDEF ) THEN
          RVAR(I,LON) = RUNDEF
        ELSE
          RVAR(I,LON) = FLOAT(IVAR(I,LON)) * 0.5
        ENDIF
102     CONTINUE
      DO 110 I=20,22
        RVAR(I,LON) = PRETAB(IVAR(I,LON))
110     CONTINUE
      RVAR(23,LON) = TMPTAB(IVAR(23,LON))
      RVAR(24,LON) = TMPVAR(IVAR(24,LON))
      RVAR(25,LON) = TMPVAR(IVAR(25,LON))
      DO 130 I=26,31
        RVAR(I,LON) = TAUTAB(IVAR(I,LON))
130     CONTINUE
      DO 135 I=29,31
        IF (RVAR(I,LON).GE.0.) RVAR(I,LON) = RVAR(I,LON) * PATHW
135     CONTINUE
    I = 32
    DO 150 ITYP=1,3
      IF ( IVAR(I,LON) .EQ. IUNDEF ) THEN
        RVAR(I,LON) = RUNDEF
      ELSE
        RVAR(I,LON) = IVAR(I,LON) * 0.5
      END IF
      RVAR(I+1,LON) = PRETAB(IVAR(I+1,LON))
      RVAR(I+2,LON) = TMPTAB(IVAR(I+2,LON))
      I = I + 3
150 CONTINUE
    DO 160 ITYP=1,15
      IF ( IVAR(I,LON) .EQ. IUNDEF ) THEN
        RVAR(I,LON) = RUNDEF
      ELSE
        RVAR(I,LON) = IVAR(I,LON) * 0.5
      END IF
      RVAR(I+1,LON) = PRETAB(IVAR(I+1,LON))
      RVAR(I+2,LON) = TMPTAB(IVAR(I+2,LON))
      RVAR(I+3,LON) = TAUTAB(IVAR(I+3,LON))
      RVAR(I+4,LON) = TAUTAB(IVAR(I+4,LON))
      IF (RVAR(I+4,LON).GE.0.0) RVAR(I+4,LON) = RVAR(I+4,LON) * PATHW
      I = I + 5
160 CONTINUE
    RVAR(116,LON) = TMPTAB(IVAR(116,LON))
    RVAR(117,LON) = TMPVAR(IVAR(117,LON))
    RVAR(118,LON) = RFLTAB(IVAR(118,LON))
    IF ( IVAR(119,LON) .EQ. IUNDEF ) THEN
      RVAR(119,LON) = RUNDEF
    ELSE
      RVAR(119,LON) = IVAR(119,LON)
    END IF
    RVAR(120,LON) = PRETAB(IVAR(120,LON))
    RVAR(121,LON) = TMPTAB(IVAR(121,LON))
    DO 190 I = 122,124
```

```
      RVAR(I,LON) = TMPTAB(IVAR(I,LON))
  190 CONTINUE
      RVAR(125,LON) = PRETAB(IVAR(125,LON))
      RVAR(126,LON) = TMPTAB(IVAR(126,LON))
      RVAR(127,LON) = TMPTAB(IVAR(127,LON))
      DO 200 I=128,129
        RVAR(I,LON) = PRWTAB(IVAR(I,LON))
  200 CONTINUE
      RVAR(130,LON) = OZNTAB(IVAR(130,LON))
  500 CONTINUE
      RETURN
      END
      SUBROUTINE MIDPRS(LON)
      PARAMETER     ( MAXVAR = 130 )
      PARAMETER     ( MAXLON = 144 )
      PARAMETER     ( RUNDEF = -1000.0 )
      COMMON /D2DATA/ LAT,NLON,IVAR(MAXVAR,MAXLON),RVAR(MAXVAR,MAXLON)
      PARAMETER     ( NLAYER = 7 )
      PARAMETER     ( NBOUND = NLAYER + 1 )
      REAL*4          PRSMID(NLAYER)
      REAL*4          PBOUND(NBOUND)
     $               /1000.,800.,680.,560.,440.,310.,180.,30./
      PSURF = RVAR(120,LON)
      PTROP = RVAR(125,LON)
      LYRSRF = 1
      LYRTRP = NLAYER
      DO 10 IBOUND=2,NBOUND
        IF ( PSURF .LE. PBOUND(IBOUND) ) LYRSRF = IBOUND
        IF ( PTROP .GT. PBOUND(IBOUND) ) LYRTRP = IBOUND-1
   10 CONTINUE
      DO 20 ILAYER=1,NLAYER
        IF ( ILAYER .LT. LYRSRF .OR. ILAYER .GT. LYRTRP ) THEN
          PRSMID(ILAYER) = RUNDEF
        ELSE IF ( ILAYER .EQ. LYRSRF ) THEN
          PRSMID(ILAYER) = ( PSURF + PBOUND(ILAYER+1) ) * 0.5
        ELSE IF ( ILAYER .EQ. LYRTRP ) THEN
          PRSMID(ILAYER) = ( PTROP + PBOUND(ILAYER) ) * 0.5
        ELSE
          PRSMID(ILAYER) = ( PBOUND(ILAYER) + PBOUND(ILAYER+1) ) * 0.5
        END IF
   20 CONTINUE
      PRINT 90,PRSMID
   90 FORMAT(/1X,'MIDPRS:  ACTUAL PRESSURE LAYER MID-POINTS (MB)',7F8.2)
      RETURN
      END
      SUBROUTINE RDANC(LUNANC,IRC)
      PARAMETER     ( MAXBOX = 6596 )
      COMMON /SQUARE/ LONLIM(2,MAXBOX)
      CHARACTER*80    HEADER
      OPEN(LUNANC,ACCESS='DIRECT',RECL=80,FORM='FORMATTED',IOSTAT=IRC)
      IF ( IRC .NE. 0 ) RETURN
      READ(LUNANC,REC=1,FMT='(A80)') HEADER
      READ(LUNANC,REC=2,FMT='(A80)') HEADER
      DO 100 IREC=1,MAXBOX
        READ(LUNANC,REC=IREC+2,FMT=110) IBOX,J,I,LONBEG,LONEND,
     $      CENLAT,CENLON,IAREA,LANDFR,ITOPOG,IVEG
        LONLIM(1,IBOX) = LONBEG
        LONLIM(2,IBOX) = LONEND
```

```
100 CONTINUE
110 FORMAT(5I4,2F9.2,I8,I6,I7,I4)
   RETURN
   END
   SUBROUTINE PRINTI(LON)
   PARAMETER ( MAXVAR = 130 )
   PARAMETER ( MAXLON = 144 )
   COMMON /D2DATA/ LAT,NLON,IVAR(MAXVAR,MAXLON),RVAR(MAXVAR,MAXLON)
   PRINT 140
140 FORMAT(/1X,'PRINTI:  COUNT VALUES FOR ALL VARIABLES')
   PRINT 145,(K,K=1,10)
145 FORMAT(1X,18X,10I8)
   DO 150 I=1,MAXVAR,10
     IEND = I + 9
     IF ( IEND .GT. MAXVAR ) IEND = MAXVAR
     PRINT 155,I,IEND,(IVAR(K,LON),K=I,IEND)
150 CONTINUE
155 FORMAT(1X,'VARIABLE (',I3.3,'-',I3.3,')',10I8)
   RETURN
   END
   SUBROUTINE PRINTR(LON)
   PARAMETER ( MAXVAR = 130 )
   PARAMETER ( MAXLON = 144 )
   COMMON /D2DATA/ LAT,NLON,IVAR(MAXVAR,MAXLON),RVAR(MAXVAR,MAXLON)
   DO 150 I=1,MAXVAR 10
     IEND = I + 9
     IF ( IEND .GT. MAXVAR ) IEND = MAXVAR
     PRINT 155,(RVAR(K,LON),K=I,IEND)
150 CONTINUE
155 FORMAT(1X,10F10.2)
   RETURN
   END
   SUBROUTINE CENTER(LON)
   PARAMETER    ( DLAT = 2.5 )
   PARAMETER    ( MAXLAT = 72 )
   COMMON /D2GRID/ NCELLS(MAXLAT),ICELLS(MAXLAT)
   PARAMETER    ( MAXVAR = 130 )
   PARAMETER    ( MAXLON = 144 )
   COMMON /D2DATA/ LAT,NLON,IVAR(MAXVAR,MAXLON),RVAR(MAXVAR,MAXLON)
   DLON = 360.0 / NLON
   CENLAT = ( LAT - 1 ) * DLAT + DLAT/2.0 - 90.0
   CENLON = ( LON - 1 ) * DLON + DLON/2.0
   PRINT 300,CENLON,CENLAT
300 FORMAT(/1X,'CENTER:  CENTER LON/LAT',2F8.2)
   RETURN
   END
   SUBROUTINE CLDHGT(LON)
   PARAMETER    ( RLAPSE = 6.5 )
   PARAMETER    ( MAXVAR = 130 )
   PARAMETER    ( MAXLON = 144 )
   COMMON /D2DATA/ LAT,NLON,IVAR(MAXVAR,MAXLON),RVAR(MAXVAR,MAXLON)

   TS = RVAR(116,LON)
   TC = RVAR(23,LON)
   HGT = ( TS - TC ) / RLAPSE * 1000.0
   PRINT 340,HGT
340 FORMAT(/1X,'CLDHGT:  CLOUD TOP HEIGHT (M)',F10.0)
```

```
      RETURN
      END
      BLOCK DATA
      PARAMETER    ( MAXCNT = 255 )
      COMMON/CNTTAB/TMPTAB(0:MAXCNT),TMPVAR(0:MAXCNT),PRETAB(0:MAXCNT),
     1          RFLTAB(0:MAXCNT),TAUTAB(0:MAXCNT),PRWTAB(0:MAXCNT),
     2          OZNTAB(0:MAXCNT)
      PARAMETER    ( MAXLAT = 72 )
      COMMON /D2GRID/ NCELLS(MAXLAT),ICELLS(MAXLAT)
      END
```

Las rutinas creadas en MATLAB, que se muestran a continuación, permiten editar los datos seleccionados D2 del *ISCCP*, según el área de trabajo durante el periodo de julio/1987 a diciembre/2004.

- **data_edit.**

```
%Routine for read & edit the data D2 saved in format x0#.txt.
clear all; close all; clc;
k=258; %k: number of files
for i=1:k
   eval(['load x0' num2str(i) '.txt;']);
   eval(['A=x0' num2str(i) ';']);
   [m,n]=size(A);
   B(i,:)=(reshape(A',m*n,1))';
end
%Separating the stations for the area of work
k=130; % k number of features
for i=1:56  % 56 all stations selected.
   eval(['C' num2str(i) '=B(:,(1+(i-1)*k):i*k);']);
end
%stratocumulus amount
SCA=[C6(:,46) C7(:,46) C8(:,46) C9(:,46) C10(:,46) C11(:,46) C12(:,46) C13(:,46) C14(:,46)...
   C18(:,46) C19(:,46) C20(:,46) C21(:,46) C22(:,46) C23(:,46) C24(:,46) C25(:,46) C26(:,46)...
   C31(:,46) C32(:,46) C33(:,46) C34(:,46) C35(:,46) C36(:,46) C37(:,46) C38(:,46) C39(:,46)...
   C43(:,46) C44(:,46) C45(:,46) C46(:,46) C47(:,46) C48(:,46) C49(:,46) C50(:,46) C51(:,46)];
a1=(mean(SCA'))';
%stratocumulus Top Pressure
SCTP=[C6(:,47) C7(:,47) C8(:,47) C9(:,47) C10(:,47) C11(:,47) C12(:,47) C13(:,47) C14(:,47)...
   C18(:,47) C19(:,47) C20(:,47) C21(:,47) C22(:,47) C23(:,47) C24(:,47) C25(:,47) C26(:,47)...
   C31(:,47) C32(:,47) C33(:,47) C34(:,47) C35(:,47) C36(:,47) C37(:,47) C38(:,47) C39(:,47)...
   C43(:,47) C44(:,47) C45(:,47) C46(:,47) C47(:,47) C48(:,47) C49(:,47) C50(:,47) C51(:,47)];
```

```
a2=(mean(SCTP'))';
%stratocumulus Top Temperature
SCTT=[C6(:,48) C7(:,48) C8(:,48) C9(:,48) C10(:,48) C11(:,48) C12(:,48) C13(:,48) C14(:,48)...
    C18(:,48) C19(:,48) C20(:,48) C21(:,48) C22(:,48) C23(:,48) C24(:,48) C25(:,48) C26(:,48)...
    C31(:,48) C32(:,48) C33(:,48) C34(:,48) C35(:,48) C36(:,48) C37(:,48) C38(:,48) C39(:,48)...
    C43(:,48) C44(:,48) C45(:,48) C46(:,48) C47(:,48) C48(:,48) C49(:,48) C50(:,48) C51(:,48)];
a3=(mean(SCTT'))';
%stratocumulus Optical Thickness
SCOT=[C6(:,49) C7(:,49) C8(:,49) C9(:,49) C10(:,49) C11(:,49) C12(:,49) C13(:,49) C14(:,49)...
    C18(:,49) C19(:,49) C20(:,49) C21(:,49) C22(:,49) C23(:,49) C24(:,49) C25(:,49) C26(:,49)...
    C31(:,49) C32(:,49) C33(:,49) C34(:,49) C35(:,49) C36(:,49) C37(:,49) C38(:,49) C39(:,49)...
    C43(:,49) C44(:,49) C45(:,49) C46(:,49) C47(:,49) C48(:,49) C49(:,49) C50(:,49) C51(:,49)];
a4=(mean(SCOT'))';
%stratocumulus Water Path
SCWP=[C6(:,50) C7(:,50) C8(:,50) C9(:,50) C10(:,50) C11(:,50) C12(:,50) C13(:,50) C14(:,50)...
    C18(:,50) C19(:,50) C20(:,50) C21(:,50) C22(:,50) C23(:,50) C24(:,50) C25(:,50) C26(:,50)...
    C31(:,50) C32(:,50) C33(:,50) C34(:,50) C35(:,50) C36(:,50) C37(:,50) C38(:,50) C39(:,50)...
    C43(:,50) C44(:,50) C45(:,50) C46(:,50) C47(:,50) C48(:,50) C49(:,50) C50(:,50) C51(:,50)];
a5=(mean(SCWP'))';
%stratus amount
SA=[C6(:,51) C7(:,51) C8(:,51) C9(:,51) C10(:,51) C11(:,51) C12(:,51) C13(:,51) C14(:,51)...
  C18(:,51) C19(:,51) C20(:,51) C21(:,51) C22(:,51) C23(:,51) C24(:,51) C25(:,51) C26(:,51)...
  C31(:,51) C32(:,51) C33(:,51) C34(:,51) C35(:,51) C36(:,51) C37(:,51) C38(:,51) C39(:,51)...
  C43(:,51) C44(:,51) C45(:,51) C46(:,51) C47(:,51) C48(:,51) C49(:,51) C50(:,51) C51(:,51)];
a6=(mean(SA'))';
%stratus Top Pressure
STP=[C6(:,52) C7(:,52) C8(:,52) C9(:,52) C10(:,52) C11(:,52) C12(:,52) C13(:,52) C14(:,52)...
  C18(:,52) C19(:,52) C20(:,52) C21(:,52) C22(:,52) C23(:,52) C24(:,52) C25(:,52) C26(:,52)...
  C31(:,52) C32(:,52) C33(:,52) C34(:,52) C35(:,52) C36(:,52) C37(:,52) C38(:,52) C39(:,52)...
  C43(:,52) C44(:,52) C45(:,52) C46(:,52) C47(:,52) C48(:,52) C49(:,52) C50(:,52) C51(:,52)];
a7=(mean(STP'))';
%stratus Top Temperature
STT=[C6(:,53) C7(:,53) C8(:,53) C9(:,53) C10(:,53) C11(:,53) C12(:,53) C13(:,53) C14(:,53)...
  C18(:,53) C19(:,53) C20(:,53) C21(:,53) C22(:,53) C23(:,53) C24(:,53) C25(:,53) C26(:,53)...
  C31(:,53) C32(:,53) C33(:,53) C34(:,53) C35(:,53) C36(:,53) C37(:,53) C38(:,53) C39(:,53)...
  C43(:,53) C44(:,53) C45(:,53) C46(:,53) C47(:,53) C48(:,53) C49(:,53) C50(:,53) C51(:,53)];
a8=(mean(STT'))';
%stratus Optical Thickness
SOT=[C6(:,54) C7(:,54) C8(:,54) C9(:,54) C10(:,54) C11(:,54) C12(:,54) C13(:,54) C14(:,54)...
  C18(:,54) C19(:,54) C20(:,54) C21(:,54) C22(:,54) C23(:,54) C24(:,54) C25(:,54) C26(:,54)...
  C31(:,54) C32(:,54) C33(:,54) C34(:,54) C35(:,54) C36(:,54) C37(:,54) C38(:,54) C39(:,54)...
  C43(:,54) C44(:,54) C45(:,54) C46(:,54) C47(:,54) C48(:,54) C49(:,54) C50(:,54) C51(:,54)];
a9=(mean(SOT'))';
%stratus Water Path
SWP=[C6(:,55) C7(:,55) C8(:,55) C9(:,55) C10(:,55) C11(:,55) C12(:,55) C13(:,55) C14(:,55)...
  C18(:,55) C19(:,55) C20(:,55) C21(:,55) C22(:,55) C23(:,55) C24(:,55) C25(:,55) C26(:,55)...
  C31(:,55) C32(:,55) C33(:,55) C34(:,55) C35(:,55) C36(:,55) C37(:,55) C38(:,55) C39(:,55)...
  C43(:,55) C44(:,55) C45(:,55) C46(:,55) C47(:,55) C48(:,55) C49(:,55) C50(:,55) C51(:,55)];
a10=(mean(SWP'))';
%nimbostratus amount
NSA=[C6(:,81) C7(:,81) C8(:,81) C9(:,81) C10(:,81) C11(:,81) C12(:,81) C13(:,81) C14(:,81)...
  C18(:,81) C19(:,81) C20(:,81) C21(:,81) C22(:,81) C23(:,81) C24(:,81) C25(:,81) C26(:,81)...
  C31(:,81) C32(:,81) C33(:,81) C34(:,81) C35(:,81) C36(:,81) C37(:,81) C38(:,81) C39(:,81)...
  C43(:,81) C44(:,81) C45(:,81) C46(:,81) C47(:,81) C48(:,81) C49(:,81) C50(:,81) C51(:,81)];
a11=(mean(NSA'))';
%nimbostratus Top Pressure
NSTP=[C6(:,82) C7(:,82) C8(:,82) C9(:,82) C10(:,82) C11(:,82) C12(:,82) C13(:,82) C14(:,82)...
  C18(:,82) C19(:,82) C20(:,82) C21(:,82) C22(:,82) C23(:,82) C24(:,82) C25(:,82) C26(:,82)...
```

```
   C31(:,82) C32(:,82) C33(:,82) C34(:,82) C35(:,82) C36(:,82) C37(:,82) C38(:,82) C39(:,82)...
   C43(:,82) C44(:,82) C45(:,82) C46(:,82) C47(:,82) C48(:,82) C49(:,82) C50(:,82) C51(:,82)];
a12=(mean(NSTP'))';
%nimbostratus Top Temperature
NSTT=[C6(:,83) C7(:,83) C8(:,83) C9(:,83) C10(:,83) C11(:,83) C12(:,83) C13(:,83) C14(:,83)...
   C18(:,83) C19(:,83) C20(:,83) C21(:,83) C22(:,83) C23(:,83) C24(:,83) C25(:,83) C26(:,83)...
   C31(:,83) C32(:,83) C33(:,83) C34(:,83) C35(:,83) C36(:,83) C37(:,83) C38(:,83) C39(:,83)...
   C43(:,83) C44(:,83) C45(:,83) C46(:,83) C47(:,83) C48(:,83) C49(:,83) C50(:,83) C51(:,83)];
a13=(mean(NSTT'))';
%nimbostratus Optical Thickness
NSOT=[C6(:,84) C7(:,84) C8(:,84) C9(:,84) C10(:,84) C11(:,84) C12(:,84) C13(:,84) C14(:,84)...
   C18(:,84) C19(:,84) C20(:,84) C21(:,84) C22(:,84) C23(:,84) C24(:,84) C25(:,84) C26(:,84)...
   C31(:,84) C32(:,84) C33(:,84) C34(:,84) C35(:,84) C36(:,84) C37(:,84) C38(:,84) C39(:,84)...
   C43(:,84) C44(:,84) C45(:,84) C46(:,84) C47(:,84) C48(:,84) C49(:,84) C50(:,84) C51(:,84)];
a14=(mean(NSOT'))';
%nimbostratus Water Path
NSWP=[C6(:,85) C7(:,85) C8(:,85) C9(:,85) C10(:,85) C11(:,85) C12(:,85) C13(:,85) C14(:,85)...
   C18(:,85) C19(:,85) C20(:,85) C21(:,85) C22(:,85) C23(:,85) C24(:,85) C25(:,85) C26(:,85)...
   C31(:,85) C32(:,85) C33(:,85) C34(:,85) C35(:,85) C36(:,85) C37(:,85) C38(:,85) C39(:,85)...
   C43(:,85) C44(:,85) C45(:,85) C46(:,85) C47(:,85) C48(:,85) C49(:,85) C50(:,85) C51(:,85)];
a15=(mean(NSWP'))';
%deepconvective amount
DCA=[C6(:,111) C7(:,111) C8(:,111) C9(:,111) C10(:,111) C11(:,111) C12(:,111) C13(:,111) C14(:,111)...
   C18(:,111) C19(:,111) C20(:,111) C21(:,111) C22(:,111) C23(:,111) C24(:,111) C25(:,111) C26(:,111)...
   C31(:,111) C32(:,111) C33(:,111) C34(:,111) C35(:,111) C36(:,111) C37(:,111) C38(:,111) C39(:,111)...
   C43(:,111) C44(:,111) C45(:,111) C46(:,111) C47(:,111) C48(:,111) C49(:,111) C50(:,111) C51(:,111)];
a16=(mean(DCA'))';
%deepconvective Top Pressure
DCTP=[C6(:,112) C7(:,112) C8(:,112) C9(:,112) C10(:,112) C11(:,112) C12(:,112) C13(:,112) C14(:,112)...
   C18(:,112) C19(:,112) C20(:,112) C21(:,112) C22(:,112) C23(:,112) C24(:,112) C25(:,112) C26(:,112)...
   C31(:,112) C32(:,112) C33(:,112) C34(:,112) C35(:,112) C36(:,112) C37(:,112) C38(:,112) C39(:,112)...
   C43(:,112) C44(:,112) C45(:,112) C46(:,112) C47(:,112) C48(:,112) C49(:,112) C50(:,112) C51(:,112)];
a17=(mean(DCTP'))';
%deepconvective Top Temperature
DCTT=[C6(:,113) C7(:,113) C8(:,113) C9(:,113) C10(:,113) C11(:,113) C12(:,113) C13(:,113) C14(:,113)...
   C18(:,113) C19(:,113) C20(:,113) C21(:,113) C22(:,113) C23(:,113) C24(:,113) C25(:,113) C26(:,113)...
   C31(:,113) C32(:,113) C33(:,113) C34(:,113) C35(:,113) C36(:,113) C37(:,113) C38(:,113) C39(:,113)...
   C43(:,113) C44(:,113) C45(:,113) C46(:,113) C47(:,113) C48(:,113) C49(:,113) C50(:,113) C51(:,113)];
a18=(mean(DCTT'))';
%deepconvective Optical Thickness
DCOT=[C6(:,114) C7(:,114) C8(:,114) C9(:,114) C10(:,114) C11(:,114) C12(:,114) C13(:,114) C14(:,114)...
   C18(:,114) C19(:,114) C20(:,114) C21(:,114) C22(:,114) C23(:,114) C24(:,114) C25(:,114) C26(:,114)...
   C31(:,114) C32(:,114) C33(:,114) C34(:,114) C35(:,114) C36(:,114) C37(:,114) C38(:,114) C39(:,114)...
   C43(:,114) C44(:,114) C45(:,114) C46(:,114) C47(:,114) C48(:,114) C49(:,114) C50(:,114) C51(:,114)];
a19=(mean(DCOT'))';
%deepconvective Water Path
DCWP=[C6(:,115) C7(:,115) C8(:,115) C9(:,115) C10(:,115) C11(:,115) C12(:,115) C13(:,115) C14(:,115)...
   C18(:,115) C19(:,115) C20(:,115) C21(:,115) C22(:,115) C23(:,115) C24(:,115) C25(:,115) C26(:,115)...
   C31(:,115) C32(:,115) C33(:,115) C34(:,115) C35(:,115) C36(:,115) C37(:,115) C38(:,115) C39(:,115)...
   C43(:,115) C44(:,115) C45(:,115) C46(:,115) C47(:,115) C48(:,115) C49(:,115) C50(:,115) C51(:,115)];
a20=(mean(DCWP'))';
% matrix for all clouds  time series
  TS_CLOUDS=[a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 a11 a12 a13 a14 a15 a16 a17 a18 a19 a20];
```