UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
VERITATE SOLA NOBIS IMPONETUR VIRILIS TOGA. 1948

Facultad Matemática, Física y Computación
Licenciatura en Matemática

TRABAJO DE DIPLOMA

SERIES CRONOLÓGICAS DE CONSUMO ELÉCTRICO Y DE PETRÓLEO EN VILLA CLARA. MODELOS Y PRONÓSTICOS.

Diplomante: Jorge Luis Morales Martínez.

Tutores: Dra. Gladys Casas Cardoso.
MSc. Humberto Mora Villegas.

Consultante: Dr. Ricardo Grau Ábalo

"Año 49 Aniversario de la Revolución"
Santa Clara
2007

Hago constar que el presente trabajo fue realizado en la Universidad Central "Marta Abreu" de Las Villas como parte de la culminación de los estudios de la especialidad de Licenciatura en Matemáticas, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

_____

Firma del autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____

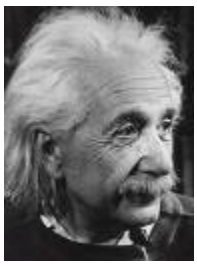
Firma del tutor *Firma del jefe del Seminario*

Dedicatoria.

A mi familia por todo el tiempo que les he robado especialmente a mi mamá Lidia Martínez Dávila por todo el amor que me mostró durante todo el tiempo que estuvo a mi lado y que siempre estará en mi corazón en todo momento, a mi papá Mateo Morales Pérez por toda su dedicación, cariño y empeño porque saliera adelante en la vida, a mi abuela, a mi prima Mayda Acosta Morales por su apreciable ayuda a lo largo de toda la carrera, a mis hermanos, a mis sobrinos, a mis compañeros de aula, a Genly León y a los bibliotecarios Miriam y Monteagudo por sus incontables ayudas y en general a todos aquellos amigos que siempre han estado a mi lado en todo momento.

Ningún resultado que obtenga podrá devolverles a ellos todo el tiempo que les he robado. Con palabras no se puede expresar el significado de familia y seres queridos

Pensamiento.



…Nunca consideres el estudio como una obligación
Sino como una oportunidad para penetrar
En el bello y maravilloso mundo del saber…

Albert Einstein.

Agradecimientos.

Con este trabajo culmina una etapa importante de nuestra vida, y el momento insta a la reflexión y en nuestra memoria se dibujan las imágenes de todos aquellos que contribuyeron de una u otra forma a la culminación exitosa del mismo, a alcanzar una meta tan deseada como esta. No quisiera mencionar sus nombres, pues cometería la grave injusticia de olvidar algunos y eso sería imperdonable. Así damos las gracias a esa inmensidad, a los que nos enseñaron poniendo en nosotros su esperanza, a aquel que un día nos dio una hoja o nos prestó un lápiz, a aquel que en un momento amargo nos hizo sonreír, al que nos escuchó, al que se mostró espontáneo, a todos aquellos que confiaron en nosotros.

También es el momento para pedir excusas por aquellas interrupciones, o por alguna tardanza o quizás porque algún día fui inoportuno.

En fin agradecer la dedicación y la paciencia, por darnos un espacio de su tiempo, un pedacito de sus vidas, porque cualquier atención, preocupación, desvelo, aunque pequeño siempre será recordado.

Especiales.

Para el MSc. Humberto Mora Villegas, Dra. Gladys Casas Cardoso, Dr.: Ricardo Grau Ábalo por sus valiosos y oportunos conocimientos, los cuales me sirvieron de gran ayuda en el desarrollo del trabajo.

Resumen.

El presente trabajo estudia el comportamiento de las series de consumo eléctrico de algunos municipios de la provincia de Villa Clara, así como las series provinciales de consumo eléctrico y de petróleo, de forma tal que se obtienen modelos matemáticos del tipo ARIMA y se realizan pronósticos a corto plazo en base a los mismos. Se constrasta con un estudio basado en un enfoque clásico de análisis de series de tiempo a través de la desestacionalización y el cálculo de la tendencia de las mismas, y para el cual se pueden obtener también modelos y pronósticos a corto plazo.

En la tesis se explica detalladamente cómo hallar los modelos del tipo ARIMA con algunos municipios representativos y los resultados fundamentales se dan a conocer en tablas resúmenes. Resulta de particular interés la incorporación de variables independientes en los modelos, ARIMA con carácter de regresores, para analizar la incidencia esperada del Período Especial en las series de consumo. Este análisis exigió nuevas consideraciones generales sobre el tratamiento de regresores en Modelos ARIMA que pueden ser útiles en estudios de cualquier índole cuando los modelos incluyen diferenciaciones regulares, estacionales o ambas.

Por último se efectúa una comparación entre los diferentes modelos para una serie en particular y entre los modelos obtenidos en dos municipios diferentes. Este análisis exigió también nuevas consideraciones teórico-prácticas sobre como distinguir de forma multivariada los coeficientes de modelos ARIMA supestamente correlacionados, entre dos modelos. Estos resultados también pueden ser útiles en estudios de cualquier índole cuando se quieren comparar, con un enfoque multivariado, dos o más modelos ARIMA para muestras de datos independientes

Todo lo anterior se realizó con la ayuda del paquete estadístico SPSS (*Statistical Packages for the Social Sciesnces*) Versión 11 para Windows.(Gupta 1999)

Introducción.

Podríamos comenzar preguntándonos donde surgió la maravillosa idea de la Revolución Energética. Para ello habría que recordar las serias dificultades enfrentadas por el Sistema Nacional en el 2004, que se analizaron en detalle en las mesas redondas de septiembre de ese año y que conllevaron, después de un estudio profundo de la situación y a partir de las experiencias del enfrentamiento a fuertes huracanes, a la puesta en práctica de nuevas concepciones para el desarrollo de un sistema electroenergético nacional más eficiente y seguro. Para poder llevar a cabo tan colosal tarea había que tomar importantes medidas para la transformación del sistema, dentro de las cuales podemos encontrar:

1. Adquisición e instalación de equipos de generación más eficientes y seguros con grupos electrógenos y motores convenientemente ubicados en distintos puntos del país.
2. Rehabilitación total de las redes de distribuciones anticuadas e ineficientes que afectaban el costo y la calidad del fluido eléctrico.
3. Priorización de los recursos mínimos necesarios para una mejor disponibilidad de las plantas del sistema electroenergético y su paso a conservación.

Todas estas medidas que exigirían una nueva concepción de generación y distribución, traerían las siguientes ventajas.

1. Valores mínimos de consumo de combustible por kilowatt/hora generado.
2. Distribución geográfica adecuada, lo cual contribuye a la protección del servicio eléctrico de la población y los objetivos económicos y sociales ante huracanes y averías.
3. Disponibilidad mayor de un 90% y muy por encima del 60% de las plantas termoeléctricas en nuestro actual sistema.
4. Valores de potencia unitaria cuya capacidad, en caso de averías, no tiene impacto significativo en la disponibilidad del sistema.

Sólo para citar algunas cifras de lo que se ha venido haciendo respecto a lo antes mencionado se puede decir que han arribado al país 6301 grupos con ese destino y de ellos están instalados 3798. Mediante los mismos se garantiza la protección entre otros objetivos 255 hospitales, 348 policlínicos, 110 clínicas estomatológicas, 245 bancos de sangre,

hogares de ancianos y de impedidos físicos y mentales y farmacias principales, 639 panaderías, 356 centros de producción, conservación y elaboración de alimentos, 37 frigoríficos, 163 centros educacionales importantes, entre otros objetivos fundamentales.

Solo después de conocer todo esto se podrá comprender mejor la Revolución Energética con la cual se posibilitará un considerable ahorro del país en divisas convertibles, un combustible noble, seguro y sano como es el caso del combustible eléctrico, sin las odiosas molestias que en todos los sentidos ocasionan los apagones que en una etapa del período especial hicieron historia y que eran frecuentes e inesperados, en un sistema y una concepción de suministros eléctricos anacrónicos

Por todo esto expresado y lo que está por venir es que nuestro Comandante en Jefe Fidel Castro Ruz expresó.

"Habrá un antes y un después de la Revolución Energética"

Como ha manifestado Fidel, lo hecho se puede considerar apenas el comienzo. Todavía la población y sobre todo el sector estatal, pueden hacer mucho por el ahorro. "El camino de la Revolución Energética constituye un proceso de aprendizaje no exento de errores por rectificar, cuyas inversiones recién empiezan. Los resultados aún resultan imposibles de medir en su total magnitud, aunque ya hay sobradas y alentadoras pruebas de su valía para la economía y el pueblo de Cuba."(Granma 2007).

Después de haber brindado una panorámica acerca de las condiciones y el por qué nace la Revolución Energética se plantea el siguiente problema de investigación.

PROBLEMA

¿Es posible con la información existente en la Oficina Nacional de Estadística de Villa Clara lograr modelos matemáticos para las series de tiempo de algunos indicadores de consumo energético y en base a los mismos, hacer pronósticos confiables a corto plazo a nivel provincial y en algunos municipios?

Esta pregunta tiene sentido porque entre otros factores, algunos indicadores dependen en alguna medida de elementos subjetivos, por ejemplo, el consumo de energía eléctrica y petróleo se planifica y controla por el hombre, o bien el problema de que existen indicadores que a lo largo de los años no siempre han sido contabilizados de la misma forma como es el caso de la productividad del trabajo. Por tanto aunque exista la información, la misma no siempre refleja la realidad.

Esta incursión en el trabajo de modelar matemáticamente el comportamiento de determinados rubros dentro de la provincia de Villa Clara, estará solamente dirigida al consumo de petróleo y energía eléctrica tanto a nivel provincial como en algunos municipios. La elección de los mismos se explica porque estas series fueron las que mejor organizadas, de mayor longitud y de mayor antigüedad existían en la Oficina Nacional de Estadística Villa Clara y además son los municipios que mayor peso presentan dentro de nuestra provincia en cuanto al renglón económico. La importancia social de estas series es obvia si se considera tan solo el impacto de estos rubros en el transporte público y en el desarrollo de la economía a partir del ahorro energético.

Por tanto en el presente trabajo se presenta el siguiente objetivo general:

OBJETIVO GENERAL

Modelar matemáticamente las series de consumo de petróleo y de energía eléctrica disponibles en la Oficina Nacional de Estadística de Villa Clara, a través de técnicas clásicas y de modelos ARIMA, con vistas a hacer pronósticos confiables de dicho consumo a corto plazo.

OBJETIVOS ESPECIFICOS

1) Modelar las series de consumo eléctrico de algunos municipios de la provincia de Villa Clara y realizar pronósticos a corto plazo en base a los mismos.
2) Mostrar cómo se puede modelar el impacto negativo económico que representó el período especial para el territorio de la provincia de Villa Clara, con énfasis en la forma de inclusión de los factores de intervención en los modelos.
3) Comparar los modelos obtenidos desde un punto de vista clásico, con los obtenidos por la modelación ARIMA y entre éstos últimos establecer comparaciones para determinar los mejores.

HIPÓTESIS

Los Modelos ARIMA son capaces de describir el comportamiento de las series de consumo de petróleo y de energía eléctrica en la provincia de Villa Clara y en algunos municipios representativos, de forma naturalmente mejor que otros modelos clásicos de series temporales. Además permiten el análisis de sistemas de influencias, como el del Período Especial, y la comparación entre municipios, de una forma eficiente.

PREGUNTAS DE INVESTIGACION

1) ¿Es posible lograr modelos ARIMA para las series de consumo eléctrico de algunos municipios de la provincia de Villa Clara con capacidad pronóstica?
2) ¿Cuál es la forma óptima de incluir en tales modelos ARIMA el posible efecto del período especial y su posible recuperación? ¿Cómo se instrumentan los regresores correspondientes para reflejar la significación de las incidencias temporales?
3) ¿Son estos modelos ARIMA superiores a los modelos clásicos de series cronológicas basados en descomposición de tendencias?
4) ¿Cómo deben compararse modelos ARIMA con las características detectadas, para diferentes series de consumo, entre municipios diferentes?

ESTRUCTURA DE LA TESIS

El capítulo 1 está dedicado al marco teórico donde se brindan conceptos generales acerca de la teoría de las series de tiempo. Además se encuentran aspectos básicos que fundamentan la metodología Box-Jenkins. El desarrollo teórico de algunos resultados no sólo permite comprender mejor el fundamento, sino que dan más claridad para la aplicación práctica, así como otros aspectos de interés que estarán presentes en alguna parte de este trabajo, en particular el uso de los regresores.

En el capítulo 2 se hace una incursión, primero teórica, y luego práctica, dentro del enfoque clásico del trabajo con series cronológicas, para hallar modelos matemáticos simples y realizar pronósticos según esta concepción. Se evidencia la insuficiencia de este enfoque para resolver el problema planteado

Por último, en el capitulo 3 se muestra mediante ejemplos, el proceso de cómo hallar los modelos del tipo ARIMA, así como una comparación entre modelos hallados según la concepción clásica y según la metodología Box-Jenkins. Quedan destacados aquí los resutltados no tan tradicionales de la forma de inclusión de los regresores. Los modelos matemáticos y los pronósticos, se dan a conocer en tablas resúmenes. El capítulo termina mostrando una comparación de la serie de mayor interés entre los modelos de dos municipios diferentes.

Capitulo 1 Análisis univariante de series temporales.

1.1 Introducción.

El análisis de las series de tiempo se aplica en muchos campos. En economía, por ejemplo, se utilizan las series de tiempo en el control de la calidad, para estudiar índices de precios, desempleo, producto nacional bruto, población. En ciencias naturales se usan para estudiar el nivel de agua en un río o presa,(Monteagudo 2007), los parámetros meteorológicos (Osés Rodríguez 2004), las medidas de poblaciones naturales (vegetales o animales). En biología las series surgen naturalmente en modelos de crecimiento. En epidemiología juegan un papel fundamental en la vigilancia epidemiológica y el estudio cronológico de factores de riesgo(Gladys Casas 1999). En las ciencias sociales representan un campo entero en si mismo. Este capítulo brinda elementos de como construir un modelo para explicar la estructura y prever la evolución de una variable que se observa a lo largo del tiempo. La variable de interés puede ser macroeconómica (índice de precios al consumo, demanda de electricidad, series de exportaciones, etc.), microeconómica (ventas de una empresa, existencias en un almacén) o de otra índole.

Se supondrá en adelante que se dispone de datos en intervalos regulares de tiempo (horas, días, meses, trimestres, años….) y se desea utilizar la posible «inercia» en el comportamiento de la serie para prever su evolución futura. Este tipo de análisis se denomina univariante o univariado porque utiliza como única información esencial la propia historia de la serie, basándose en la hipótesis central de que las condiciones futuras serán análogas a las pasadas. Los modelos univariantes son especialmente útiles para la previsión a corto plazo; ellos pueden llegar a ser útiles en un pronóstico a mediano plazo, sobre todo si se tienen comportamientos periódicos en la serie; pero de manera general se puede decir que para el pronóstico a mediano y largo plazo respectivamente se obtienen mejores resultados si se tiene en cuenta la evolución de otras variables, relacionándolas con la que nos interesa prever mediante un modelo de regresión dinámica o modelo de transferencia (modelos “multivariados” o “multivariantes”). Los conceptos básicos necesarios para el análisis de series se presentan en las secciones que se tratarán a continuación, en particular aquellos que estudian la familia de procesos básica para el análisis univariante de series temporales: los modelos ARIMA.

1.2 Procesos estocásticos y series temporales.

En este epígrafe se explican los conceptos fundamentales de procesos estocásticos y series de tiempo.

1.2.1 El concepto de proceso estocástico.

Un proceso estocástico $Z(t)$ es una función que a cada instante de tiempo t le hace corresponder una variable aleatoria Z_t.

Los procesos estocásticos son modelos matemáticos bastantes bondadosos para poder describir el comportamiento de los fenómenos aleatorios que se desarrollan en el tiempo. Por otra parte para poder caracterizar de manera general el proceso estocástico se requiere una función de distribución multivariada, la cual por lo general no se conoce en la mayoría de los casos. No obstante, se acude a la experimentación, a la observación, para obtener datos que permitan hacer inferencias acerca del proceso estocástico estudiado.

Los datos recogidos de la observación se organizan cronológicamente, y así se obtienen las llamadas series cronológicas o series temporales.

Una serie temporal $z(t)$ es un conjunto de observaciones secuencialmente generadas en el tiempo, de modo que le corresponda un valor z_t a cada instante t observado.

En nuestro caso particularmente nos interesa analizar el caso en que los valores de la serie estén influenciados por factores aleatorios. Por ello, formalmente hablando una serie temporal puede ser considerada como una colección de variables aleatorias $\{ Z_t, \ t \in T \}$ donde T es un conjunto de índices, por lo general el conjunto de los números naturales.

Así, los valores de la serie pueden ser vistos como salidas de un proceso estocástico, de tal manera que cada valor z_t de la serie temporal puede ser considerado como una observación de una de las variables aleatorias Z_t que integran el proceso y la serie temporal de n observaciones sucesivas $(z_1(t), z_2(t), ..., z_n(t))$ puede ser considerada como una muestra de una población de series temporales $(Z_1(t), Z_2(t), ..., Z_n(t))$ que podían haber sido generadas por el proceso estocástico estudiado. Por esta sencilla razón, en este trabajo se utiliza indistintamente el concepto de proceso o el de serie como lo hacen otros autores, por ejemplo (Peña 1999).

Se nombra **función de medias** del proceso a una función del tiempo que proporciona las medias de las distribuciones marginales Z_t para cada instante:

$$\mu_t = E\left[z_t\right] \qquad (1.1.)$$

Un caso particular importante es aquel en que todas las variables tengan la misma media; es decir, la función de medias es constante. Entonces, las realizaciones del proceso no mostrarán tendencia creciente o decreciente y diremos que el proceso es estable en la media. Si por el contrario, las medias crecen con el tiempo, las observaciones en distintos momentos mostrarán esta tendencia. La figura 1(a) presenta una serie temporal que ha sido generada simulando un proceso estable en la media, mientras que la 1(b) proviene de otro con media no estable.

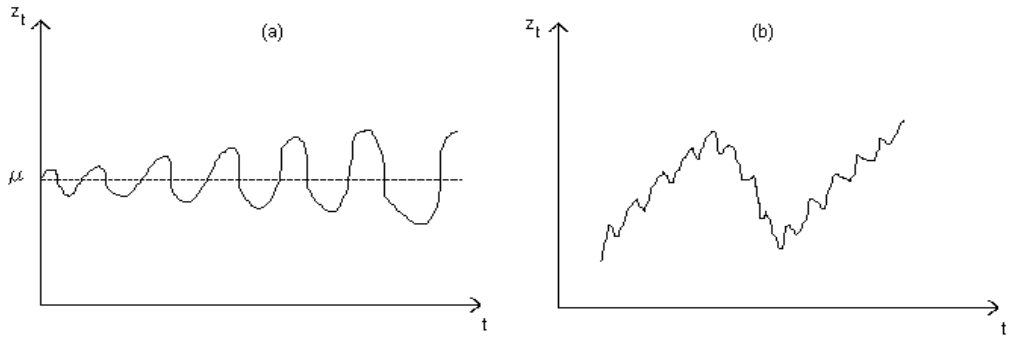


Figura 1. Procesos con media estable (caso a) y no estable (caso b).

La **función de varianzas** del proceso proporciona las varianzas en cada instante temporal:

$$\sigma_t^2 = Var\left(z_t\right) \qquad (1.2.)$$

Y se dice que el proceso es estable en la varianza si ésta es constante en el tiempo. El proceso puede ser estable en la media y no en la varianza y al revés (la serie de la figura 1(a) no es estable en la varianza, mientras que la 1(b) si parece serlo.

La estructura de dependencia lineal entre las variables aleatorias del proceso se representa por las funciones de covarianza y correlación. Se nombra **función de autocovarianzas** del proceso, a la función que describe las covarianzas en dos instantes cualesquiera:

$$Cov\left(t,t+j\right) = Cov\left(z_t, z_{t+j}\right) = E\left[\left(z_t - \mu_t\right)\left(z_{t+j} - \mu_{t+j}\right)\right] \qquad (1.3.)$$

Y **función de autocorrelación** a la estandarización de la función de covarianzas

$$\rho_{(t,t+j)} = \frac{Cov(t,t+j)}{\sigma t \sigma t+j} \qquad (1.4.)$$

En general estas dos funciones dependen de dos parámetros (t, j), siendo t el instante inicial y j el intervalo entre observaciones. Una condición de estabilidad que aparece en muchos fenómenos dinámicos, es que la dependencia entre dos observaciones solo depende del intervalo entre ellas y no del origen considerado. Entonces se puede escribir:

$Cov(t_1, t_{1+k}) = Cov(t_2, t_{2+k}) = \gamma_k$ **k= 0,1,2,3,4…**

y las autocovarianzas dependen exclusivamente de la distancia entre las variables. En este caso la relación entre z_t y z_{t-k}, es siempre igual a la relación entre z_t y z_{t+k}.

Si se estudia a lo largo del tiempo la serie de ventas de una empresa o la evolución del ozono en la atmósfera, el proceso estocástico existe conceptualmente (admitiendo que cada dato es un valor particular de todos los que podrían haberse observado en dicho instante y que definen la distribución de la variable Z_t) pero solo se dispone de un valor de cada variable en cada instante. Para poder estimar las características **transversales** del proceso (medias, varianzas, etc.) a partir de su evolución **longitudinal** es necesario suponer que las propiedades **transversales** (distribución de la variable en cada instante) son estables a lo largo del tiempo. Esto conduce al concepto de estacionariedad que se define a continuación.

1.2.2 Series estacionarias.

Se dice que una serie temporal es **estacionaria en sentido débil,** si existen y son estables la media, la varianza y las covarianzas, es decir, si para toda t:

$(i)\mu_t = \mu = cte.$

$(ii)\sigma^2{}_t = \sigma^2 = cte.$

$(iii)Cov(t, t+k) = Cov(t, t-k) = \gamma_k \qquad k = 0,1,2,\ldots$

Teniendo en cuenta que para una serie estacionaria $\gamma_0 = \sigma^2$ la función de autocorrelación con retardo K, $K \geq 0$ se calcula mediante:

$$\boxed{\rho_k = \frac{\gamma_k}{\gamma_0}}$$

Se denomina **función de autocorrelación simple (fas**), en inglés ACF, acrónimo de autocorrelation function, o simplemente correlograma, a la representación de los coeficientes de autocorrelación en función del retardo K, $K \geq 0$. La figura 2 presenta el correlograma de dos series estacionarias.

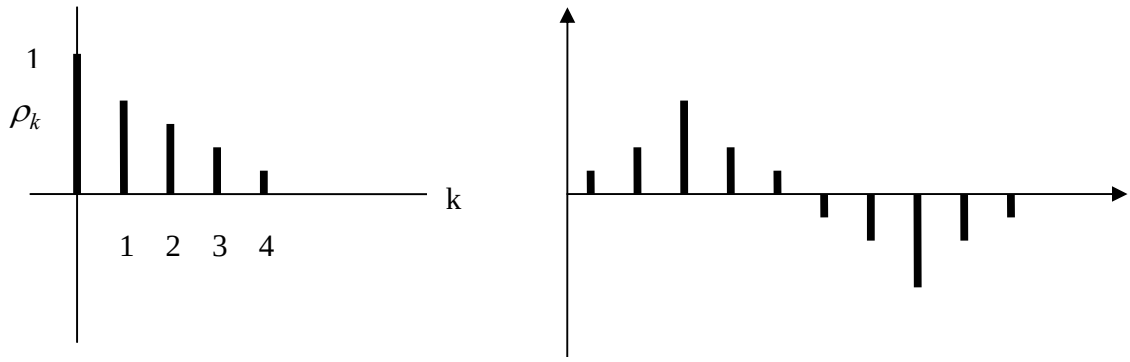


Figura 2. Correlograma de dos procesos estacionarios

En el primero de ellos la dependencia entre observaciones tiende a cero al aumentar el retardo y se dice que el proceso es *ergódico,* mientras que el segundo no lo es. La ergodicidad es necesaria para poder estimar las características del proceso a partir de una única realización, ya que si el proceso no es ergódico al aumentar el tamaño muestral no se adquiere información adicional por ser todas las observaciones muy dependientes entre sí, y no será posible obtener estimadores consistentes de los parámetros. En adelante se supondrá siempre procesos estacionarios ergódicos.

1.2.3 Proceso de ruido blanco.

Una serie estacionaria muy importante es la definida por:

$1) E[z_t] = 0$

$2) Var(z_t) = \sigma^2$

$3) Cov(z_t, z_{t-k}) = 0, \qquad k = 0,1,2,\ldots$

Este proceso se denomina *proceso de ruido blanco.* En este tipo de proceso, conocer los valores pasados no proporciona ninguna información sobre el futuro ya que el proceso "no tiene memoria". Si se supone adicionalmente que en un proceso de ruido todas las variables tienen distribución normal, la no correlación garantiza la independencia. De forma general, se denotará el ruido blanco por e_t

1.2.4 Series Integradas.

La gran mayoría de las series que observamos no son estacionarias, y su nivel medio varía con el tiempo. Sin embargo, es frecuente que la serie se convierta en estacionaria al diferenciarla. Se nombra serie «primera diferencia» del $\{z_t\}$, $t = 1,.....n.$ a una nueva serie $\{\omega_t\}$ obtenida mediante: $\omega_t = z_t - z_{t-1}$

Análogamente, llamaremos serie «segunda diferencia» de la original a:

$$Y_t = \omega_t - \omega_{t-1} = z_t - 2z_{t-1} + z_{t-2}.$$

Se dice que una serie es integrada de orden h cuando al diferenciarla h veces se obtiene una serie estacionaria. Obsérvese que series que sean la suma de una tendencia polinómica y una serie estacionaria serán integradas. Por ejemplo, la serie:

$z_t = a + bt + u_t$ donde u_t es estacionaria, es integrada de orden uno o I(1).

En general, series generadas como una suma de una tendencia polinómica de orden h más una serie estacionaria cualquiera sea u_t, serán integradas de orden h.

1.3 Procesos autorregresivos.

Una clase muy importante de series estacionarias son los procesos autorregresivos, que resultan de imponer una dependencia lineal entre las variables de la serie, similar a la de una ecuación de regresión.

1.3.1 Serie autorregresiva de primer orden AR(1).

La forma de dependencia más simple es relacionar z_t con z_{t-1} linealmente mediante la ecuación de «autoregresión»:

$$Z_t = c + \phi Z_{t-1} + e_t \qquad (1.5.)$$

Donde c y ϕ son constantes a determinar y e_t es un proceso de ruido blanco, (es decir, análogo a la perturbación en un modelo de regresión: distribución normal con $E[\,e_t\,]=0$,

Var [e_t]= σ^2 , y Cov (e_t , e_{t+k})=0 independiente de z_{t-k} para todo k positivo. El proceso (1.5) se denomina proceso autorregresivo de primer orden, AR(1).Estos procesos presentan ciertas propiedades muy importantes las cuales las podemos apreciar en (Mora Villegas 2003).

1.3.2 Serie autorregresiva general AR (p).

Se dice que una serie z_t es autorregresiva de orden p si

$\bar{Z} = \phi_1 \overline{Z}_{t-1} + ...\phi_p \overline{Z}_{t-p} + e_t$ donde $\overline{z}_t = z_t - \mu$,(μ es la media de la constante de la serie y e_t es un ruido blanco independiente de Z_{t-h}) para todo $h \geq 0$.

Para facilitar el manejo de estos y otros procesos, se define el operador de retardo, B, por

$BZ_t = Z_{t-1}$

$B^2 Z_2 = BBZ_t = Z_{t-2}$

.....

.....

$B^k Z_k = B \underset{K}{.....} BZ_t = Z_{t-k}$

Al aplicar este operador k veces a una constante, esta no se modifica para k=0 y se anula para $k \geq 1$.Con esta notación de operadores, la ecuación de un AR(p) es:

$\bar{Z} = (\phi_1 B + ...\phi_p B^p)\bar{Z}_t + e_t$ es decir

$(1 - \phi_1 B - \phi_2 B^2 ..... - \phi_p B^p)\bar{Z}_t = e_t$

Y llamando $\phi_p(B) = 1 - \phi_1 B - ..... - \phi_p B^p$ al polinomio de grado p en el operador de retardo, cuyo primer término es la unidad, se abrevia:

$\phi_p(B)\bar{z}_t = e_t$ que es la expresión general de una serie autorregresiva. Se nombra ecuación característica de la serie a la ecuación: $\boxed{\phi_p(B) = 0}$

considerada como función de B. En general, las condiciones que garantizan que un tal proceso es estacionario se formulan en términos de los coeficientes o las raíces de la ecuación característica, como una generalización de la condición $|\phi| < 1$ que se vió en el caso de orden 1, (ver generalización en (Grau 1996)).

1.3.3 La función de autocorrelación parcial (fap).

Determinar el orden de un proceso autorregresivo a partir de su función de autocorrelación es difícil. En general ésta función es una mezcla de decrecimientos

exponenciales y sinusoidales, que se amortiguan al avanzar el retardo, y no presenta rasgos fácilmente identificables con el orden del proceso. Por ejemplo, en un proceso autorregresivo simple de orden 1, la observación en cada tiempo *t* depende de la observación en *t*-1; pero esta a su vez, depende entonces de la observación en *t*-2, por tanto la observación en *t* depende transitivamente de la observación en *t*-2. La única forma de distinguir entonces un proceso autorregresivo de orden 1 de un proceso autorregresivo de orden 2 o superior es mediante una función que determine la autocorrelación "parcial" entre la observación en *t* y la observación en *t*-2 quitando el efecto ya explicado por la correlación con el instante intermedio *t*-1. Para resolver este problema se introduce la función de autocorrelación parcial (fap), en inglés PACF, acrónimo de partial autocorrelation function.

Se define el coeficiente de autocorrelación parcial de orden h como una medida de la relación lineal entre observaciones separadas h períodos con independencia de los valores intermedios mediante:

$$fap(h) = \frac{Cov(Z_t - \rho_1 Z_{t-1} - \rho_2 Z_{t-2} \ldots \rho_{h-1} Z_{t-h+1}, Z_{t-h} - \rho_{h-1} Z_{t-1} - \ldots - \rho_1 z_{t-h+1})}{Var(Z_t - \rho_1 Z_{t-1} - \rho_2 Z_{t-2} \ldots \rho_{h-1} Z_{t-h+1})}$$

donde ρ_i es el coeficiente de correlación i-ésimo.

Se puede deducir que un proceso AR(p) tendrá los p primeros coeficientes de autocorrelación parcial distintos de cero, y por tanto, en la fap el número de coeficientes distintos de cero indica el orden del proceso AR. Esta propiedad va a ser clave para identificar el orden de un proceso autorregresivo. **El anexo I-1** resume la fas (ACF) y la fap (PACF) de distintos procesos autorregresivos.

1.4 Series de media móvil.

Veamos ahora otro tipo de series estacionarias que en algún sentido son inversas de los procesos autorregresivos

1.4.1 Descomposición de Wold.

Las series autorregresivas estudiadas en la sección anterior son casos particulares de una representación general de procesos estacionarios obtenida por Wold (1938). Este autor demostró que todo proceso estocástico débilmente estacionario de media cero, Z_t que no

contenga componentes deterministas, puede escribirse como una función lineal de variables aleatorias incorrelacionadas e_t , como:

$$\bar{Z}_t = e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + ... = \sum_{i=0}^{\infty} \psi_i e_{t-i} \qquad (\psi_0 = 1) \tag{1.6}$$

donde $E[e_t] = 0, Var(e_t) = \sigma^2, E[e_t e_{t-k}] = 0,$ K>1. Utilizando el operador de retardo se puede escribir:

$$\bar{Z}_t = \psi(B) e_t \tag{1.7}$$

siendo $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + ...$ una serie polinómica en el operador de retardo B. Se llamará a (1.11) *la representación lineal general* de un proceso estacionario no determinista.

La representación de Wold (1.7) puede interpretarse como la descomposición de un vector $\bar{z}_t$ en sus coordenadas sobre ejes ortogonales asociados a cada elemento e_t. Suponiendo que el proceso comienza en t=0, los tres primeros valores observados forman un vector $\vec{z} = (z_1, z_2, z_3)$ que puede descomponerse en tres ejes ortogonales (a_1, a_2, a_3), (recuérdese que incorrelación equivale a ortogonalidad). Todas las posibles realizaciones del proceso de tamaño 3 pueden obtenerse generando tres valores al azar (a_1, a_2, a_3) de una distribución $N(0, \sigma^2)$ y tomando como coordenadas de la realización $(\psi_1 a_1, \psi_2 a_2, \psi_3 a_3)$. Generalizando esta idea para t variables, $(z_1, z_2, ...z_t)$, aumentará la dimensión de Z, realización del proceso, y en el límite nos moveremos en un espacio de dimensión infinita que es la representación (1.7.). Los coeficientes ψ_i tienen que verificar ciertas condiciones para que el proceso sea estacionario. La varianza de z_t en (1.6.) será:

$$Var(z_t) = \gamma_0 = \sigma^2 \sum_{i=0}^{\infty} \psi^2{}_i \tag{1.8}$$

y para que el proceso sea estacionario, la serie $\sum_{i=0}^{\infty} \psi^2{}_i$ debe ser convergente.

1.4.2 Serie de media móvil de orden uno MA (1).

Se llaman series de media móvil de orden q, MA (q), a aquellos casos particulares del proceso lineal general en los que únicamente los q primeros coeficientes ψ_i son no nulos. Estos procesos serán estacionarios ya que la condición (1.8) se cumplirá siempre.

La serie MA (1) será:

$\bar{Z}_t = e_t - \theta \cdot e_{t-1}$ que es un caso particular $(1-\phi B)^{-1}$ de (1.7) con $\psi_1 = -\theta,\ \psi_i = 0;\ i \geq 2$.

Este proceso puede escribirse:

$\bar{Z}_t = (1-\theta B)e_t$ y suponiendo $|\theta| < 1$, existe el operador inverso $(1-\phi B)^{-1}$ y tendremos: $(1+\theta B+\theta^2\ B^2+.....)\bar{Z}_t = e_t$ y por tanto, un proceso MA(1) equivale a un AR(∞) cuyos coeficientes decrecen en progresión geométrica. Este resultado es dual de otro resultado que se puede demostrar y que nos dice que un AR(1) equivale a un MA(∞).

Función de autocorrelación simple y parcial.

Para obtener la función de autocorrelación simple y parcial de un proceso MA(1) multiplicando por $\bar{z}_{t-1}$

$$\gamma_1 = -\theta\sigma^2$$

$$\gamma_2 = 0$$

además $\gamma_0 = \sigma^2(1+\theta^2)$ por tanto:

$$\rho_1 = \frac{-\theta}{1+\theta^2} \qquad \rho_k = 0 \quad k > 1$$

y la función de autocorrelación simple tendrá únicamente un valor distinto de cero en el primer retardo. Se puede observar que la función de autocorrelación simple (fas) de un proceso MA(1) es similar a la función de autocorrelación parcial (fap) de un proceso AR(1). Esta dualidad también se presenta también en la función de autocorrelación parcial, fap.

1.4.3 El proceso de media móvil general, MA(q).

Un proceso MA(q) tiene la representación general:

$$\overline{Z}_t = (1 - \theta_1 B - \theta_2 B^2 - ..... - \theta_q B^q) e_t$$

que suele escribirse:

$$\overline{Z}_t = \theta_q(B) e_t$$

La serie será inversible si las raíces de $\theta(B) = 0$ son, en módulo, mayores que la unidad. Puede demostrarse en virtud de la dualidad entre procesos AR y MA que la fap de un MA(q) tiene la estructura de la fas de un AR(q). **El anexo I-2** presenta estas funciones para algunos procesos frecuentes.

1.5 Procesos ARMA.

A continuación explicamos como se pueden combinar procesos autoregresivos (AR) con procesos de medias móviles (MA) para formar un nuevo tipo de procesos: ARMA

1.5.1 Proceso ARMA (1,1).

Las series AR y MA son aproximaciones al proceso lineal general MA(∞) desde puntos de vista complementarios: los AR suponen estructura MA(∞), pero imponen restricciones sobre las pautas de decrecimiento de los coeficientes ψ_i; los MA suponen un número finito de términos, pero en cambio no imponen restricciones sobre sus valores. Los procesos ARMA intentan combinar estas ventajas y permiten representar de forma escueta (utilizando pocos parámetros) procesos cuyos primeros q coeficientes son cualesquiera, y los siguientes decrecen según leyes simples. Matemáticamente resultan de añadir estructura MA a un proceso AR o viceversa. La serie más simple es el ARMA(1,1), que se escribe:

$$(1 - \phi_1 B)\,\bar{z}_t = (1 - \theta_1 B) e_t$$

donde $|\phi_1| < 1$ para que el proceso sea estacionario, y $|\theta| < 1$ para que sea inversible. Se verifica que la función de autocorrelación simple es:

$$\rho_1 = \phi - \theta \frac{\sigma^2}{\gamma_0}$$

$$\rho_k = \phi \rho_{k-1} \quad k > 1$$

La fas de un ARMA(1,1) tiene un decrecimiento exponencial que comienza a partir de ρ_1, y no de $\rho_0 = 1$ como en el AR(1).

Para calcular la función de autocorrelación parcial observe que el ARMA(1,1) puede escribirse:

$$(1-\theta B)^{-1}(1-\phi B)\bar{Z}_t = e_t$$

$$\bar{Z}_t = (\phi-\theta)\bar{Z}_{t-1} + \theta(\phi-\theta)\bar{Z}_{t-2} + \theta^2(\phi-\theta)\bar{Z}_{t-3} + ...e_t$$

El efecto directo de $\bar{z}_{t-k}$ sobre $\bar{z}_t$ decrece geométricamente con θ^k y por tanto la fap tendrá un decrecimiento geométrico a partir de un valor inicial.

En conclusión, en un proceso ARMA(1,1) la fas y la fap tienen una estructura similar. (**ver anexo I-3**).

1.5.2 Procesos ARMA(p,q).

El proceso ARMA(p,q) será:

$$(1-\phi_1 B - \phi_2 B^2 - ..... - \phi_p B^p)\bar{Z}_t = (1-\theta_1 B - \theta_2 B^2 - ..... - \theta_q B^q)e_t$$

o, en notación compacta, $\phi_p(B)\bar{Z}_t = \theta_q(B)e_t$

El proceso será estacionario si las raíces de $\phi_p(B) = 0$ están fuera del circulo unidad, e inversible si lo están las de $\phi_q(B) = 0$.

En esta serie se cumple en cuanto a las fas que:

(1) tendrá q-p+1 valores iniciales con cualquier estructura,

(2) decrecerá a partir del coeficiente q-p como una mezcla de exponenciales y sinusoides, determinada por la parte autorregresiva. Puede comprobarse que la fap tendrá una estructura similar. **La tabla 1(anexo I-3)** resume estas características.

1.6 Procesos no estacionarios.

Se analizarán ahora algunos modelos en los que no hay estacionariedad y que son importantes para estudiar procesos estacionarios

1.6.1 Paseo aleatorio.

Se ha visto que las series MA finitas son siempre estacionarias y que las AR lo son si las raíces de $\phi(B) = 0$ están fuera del círculo unidad.

Al considerar el AR(1):

$\bar{Z}_t = \phi \overline{Z}_{t-1} + e_t$ si $|\phi| > 1$ el proceso es explosivo. Si $|\phi| = 1$ el proceso no es estacionario, pero tampoco es explosivo, y pertenece a la clase de series integradas de orden uno (ya que su primera diferencia, $\bar{Z}_t - \bar{Z}_{t-1} = e_t$ si es un proceso estacionario). Se trata concretamente de la integración de un ruido blanco. Este proceso se denomina paseo aleatorio.

Definiendo el *operador diferencia* ∇ =1-B, el paseo aleatorio se escribe:

$$\nabla \bar{Z}_t = e_t \qquad (1.9)$$

En este proceso se verifica que $Cov(t, t+k) = \sigma^2 t$

y la función de autocorrelación es:

$\rho(t, t+k) = \frac{t}{\sqrt{t(t+k)}}$ y si t es grande, los coeficientes de la función de autocorrelación serán próximos a uno y decrecerán muy lentamente con k.

1.6.2 Procesos ARIMA.

El Proceso anterior se ha obtenido admitiendo que la raíz de la parte AR de los procesos AR(1) y ARMA(1,1) es unitaria, con lo que se convierten en no estacionarios. Esta idea puede generalizarse para cualquier proceso ARMA permitiendo una o varias raíces unitarias en el operador AR. Se obtienen entonces procesos del tipo:

$$(1 - \phi_1 B - \phi_2 B^2 - ..... - \phi_p B^p)(1-B)^d \bar{Z}_t = (1 - \theta_1 B - \theta_2 B^2 - ..... - \theta_q B^q) e_t$$

que llamaremos procesos **ARIMA(p,d,q).** En esta notación **p** es el orden de la parte autorregresiva estacionaria, **d** es el número de raíces unitarias (orden de homogeneidad del proceso) y **q** es el orden de la parte de la media móvil. Llamando $\nabla = 1 - B$ al operador diferencia, el proceso anterior suele escribirse:

$$ARIMA(p,d,q) \quad \phi_p(B) \nabla^d \bar{Z}_t = \theta_q(B) e_t \qquad (1.10)$$

El nombre ARIMA proviene de las iniciales en inglés de los procesos autorregresivos integrados de media móvil (autoregresive integrated moving average), donde «integrado»

significa, como se indico en 1.2.4, que si llamamos $w_t = \nabla^d \bar{z}_t$ a la serie diferenciada d veces de $\bar{z}_t$, esta es inversamente la serie d veces integrada de w_t. En efecto, si

$$w_t = (1-B)\bar{z}_t$$

teniendo en cuenta que:

$$(1-B)^{-1} = 1 + B + B^2 + B^3 + ... \qquad (1.11)$$

resulta: $Z_t = (1-B)^{-1} W_t = W_t + W_{t-1} + W_{t-2} + ... = \sum_{k=0}^{\infty} W_{t-k}$

Se ha visto en las secciones anteriores un ejemplo de estas series; el paseo aleatorio es el modelo ARIMA(0,1,0). El cual se caracteriza porque la función de autocorrelación simple tiene coeficientes que decrecen lentamente. Todas las series ARIMA no estacionarias tienen esta propiedad general.

Procesos ARIMA estacionales.

El concepto de estacionalidad.

Un tipo especial de no estacionaridad presente en muchas series es la *estacionalidad,* entendiendo por ello una pauta regular de comportamiento periódico en la serie. Por ejemplo, existirá estacionalidad si los meses de enero tienden a ser similares en distintos años y lo mismo, los meses de febrero, los de marzo, etc. Estacionalidad es en cierto sentido sinónimo de periodicidad, pero como trabajamos con datos discretos, ella se materializa en el número de observaciones que se observan en un período. Así una serie que representa hipotéticamente datos continuos con periodicidad de un año, tiene una estacionalidad de 12 si las observaciones son mensuales, pero tiene una estacionalidad de 4 si las observaciones son trimestrales, o incluso de aproximadamente 365, si las observaciones son diarias. Dejando aparte otros efectos, la estacionalidad hace a la serie no estacionaria, ya que el valor medio μ_t variará de unos meses a otros. Para más información ver (Grau 1996) y (Mora Villegas 2003).

Formulación general

Considere una serie estacional z_t con período conocido. Como ejemplo supondremos que los datos son mensuales y que se dispone de h años completos con un total de n=12h observaciones. Si hay estacionalidad se puede dividir la serie total en 12 series de h datos,

una por mes, que llamaremos $y_\tau^{(1)}$....., $y_\tau^{(12)}$ con τ =1,…, h. La relación entre estas series y la primitiva, z_t, será:

$$y_\tau^{(j)} = z_{t+12(\tau-1)} \qquad (\tau = 1,.....h) \tag{1.12}$$

es decir, el índice τ indica años, el j (j=1,…..12) mes dentro del año y el t meses a partir del origen, existiendo la relación , $t = j + 12(\tau - 1)$.

Ajustemos ahora un modelo ARIMA para cada una de estas series y *supongamos que este modelo es exactamente el mismo para todas.* Este modelo será del tipo:

$$\left(1-\Phi_1 B-...\Phi_P B^P\right)(1-B)^D y_\tau^{(j)} = \left(1-\Theta_1 B-...-\Theta_Q B^Q\right)u_\tau^{(j)} \ \tau = 1,...h \tag{1.13}$$

y forzosamente $D \geq 1$ si hay estacionalidad . En efecto, si D=0 y las series fueran estacionarias, su modelo podría escribirse:

$$y_\tau^{(j)} = \mu_j + \psi_j(B)\mu_\tau^{(j)} \qquad (\tau = 1,...h)$$

donde μ_τ es la media del mes j y $W_j(B)$ seria una serie polinómica en el operador de retardo que resulta del producto del inverso del polinomio autorregresivo en B con el polinomio media móvil en B. Si hay estacionalidad, estas medias son diferentes, con lo que las 12 series no pueden tener un modelo común.

Los modelos para las 12 series pueden escribirse conjuntamente utilizando (1.12):

$$By_\tau^{(j)} = y_{\tau-1}^{(j)} = z_{j+12(\tau-2)} = z_{j-12+12(\tau-1)} = B^{12} z_{j+12(\tau-1)}$$

es decir, aplicar B a y_τ, equivale a aplicar B^{12} a z_t.

Se define ahora una serie de ruido común, α_t, asignando a cada mes t el ruido del modelo univariante (1.13) correspondiente a dicho mes. En consecuencia, α_t se obtendrá a partir de las 12 series $\mu_t^{(j)}$ mediante:

$$u_\tau^{(j)} = \alpha_{j+12(\tau-1)} \qquad \tau = 1,....h$$

que coincide lógicamente con (1.12).

Entonces, llamando s al período estacional, los modelos (1.13) pueden escribirse:

$$\left(1-\Phi_1 B^s-...-\Phi_P B^{Ps}\right)\left(1-B^s\right)^D \overline{z}_t = \left(1-\Theta_1 B-...\Theta_Q B^{Qs}\right)a_t \tag{1.14}$$

donde $t = 1,...n$ y ahora el modelo ARIMA se formula en B^s (B^{12} para datos mensuales).

Las series $u_\tau^{(j)}$ son, por hipótesis, ruido blanco, pero la serie conjunta $\alpha_t\,(t = 1,...,12h)$ normalmente no lo será, ya que existirá en general dependencia entre observaciones continuas. Llamaremos *estructura regular* a la asociada a los intervalos naturales de medida de la serie (meses en el ejemplo), para diferenciarla de la estructura estacional asociada a intervalos de amplitud s. Suponiendo que α_t sigue el proceso ARIMA regular:

$$\phi_p(B)\nabla^d \alpha_t = \theta_q(B)e_t \qquad (1.15)$$

Sustituyendo este *modelo regular* en el modelo estacional (1.14) se obtiene el *modelo completo* para el proceso observado:

$$\Phi_p(B^s)\phi_p(B)\nabla^d\nabla_s{}^D Z_t = \theta_q(B)\Theta_Q(B^s)e_t \qquad (1.16)$$

que se denomina modelo $ARIMA(p,d,q)\times(P,D,Q)_S$.

Estos modelos introducidos por Box y Jenkins (1976), representan de forma simple muchos fenómenos reales que se encuentran en la práctica. Observe que estos *modelos se basan en la hipótesis central de que la relación de dependencia estacional (1.13) es la misma para todas las estaciones, entendidas como elementos de un periodo, por ejemplo meses.* **El anexo I-4** presenta algunos ejemplos de la fas de modelos estacionales.

1.7 Complementos teóricos: estimación, diagnóstico, y pronósticos en modelos ARIMA.

Ya se conoce que una serie ARIMA (p,d,q)(P,D,Q)S muestra necesariamente cierto comportamiento de las funciones ACF(h) y PACF(h) que sirven para identificar el modelo. La teoría matemática de series de tiempo abarca criterios para lograr las estimaciones de máxima verosimilitud de dichas funciones a partir de datos observados o realización de una serie.

Una vez identificada la estructura ARIMA (p,d,q)(P,D,Q)S a la cual responde (probablemente) la muestra de una serie de tiempo, el paso próximo y más importante es la estimación estadística de los parámetros del modelo. La argumentación matemática de la estimación estadística de los parámetros esta descrita en detalle en el capítulo 7 del libro (Guerrero 1991) y se fundamenta en la teoría de estimadores de máxima verosimilitud. Cuando la serie no tiene valores perdidos, los estimadores iniciales se hacen sobre la base

de un criterio de máxima verosimilitud y el algoritmo resulta particularmente rápido. Se conoce así como algoritmo de Marquardt-Melard y es el que utilizan la mayor parte de los paquetes serios de análisis de series de tiempo.

El algoritmo de Marguardt-Melard es iterativo y como criterio de convergencia o de finalización del algoritmo se pueden utilizar algunos o varios de los siguientes:

- **Un valor epsilon, (por ejemplo $\varepsilon = 0.001$).** El proceso terminaría según este criterio cuando el cambio en todos los parámetros estimados fuera menor que epsilon.
- **Porcentaje de variación de la suma de los cuadrados.** El proceso iterativo debe terminar si el cambio relativo en la suma de cuadrados es menor que cierta cantidad prefijada, que se denomina"SSQ Percentage", por ejemplo, SSQ=0.001%.
- **Un valor máximo de la constante de Marquardt**. Esta es una constante que se utiliza por el algoritmo de Marquardt y que se actualiza en cada iteración. Generalmente esta constante debe ser cercana a cero o tender a este cuando se obtienen las estimaciones finales. Un valor grande de esta constante indica problemas condicionantes en los datos, por ello se formula un criterio de terminación (más bien de aborto) del algoritmo en términos de que la constante no rebase un valor máximo prefijado, por ejemplo10^9, o el analista decide rechazar un modelo si la constante muestra valores oscilantes sin una franca tendencia a cero.
- **Número máximo de iteraciones.** Si se utiliza el algoritmo de Marquartd-Melard y el modelo esta correctamente identificado se garantiza alta velocidad de convergencia. Por tanto la necesidad de muchas iteraciones puede ser un indicador de un problema y se utiliza un máximo, por ejemplo, 10 para abortar.

Es importante destacar que en la estimación de parámetros se debe perseguir varios objetivos:

1) Que los parámetros resulten significativamente diferentes de cero.
2) Que los valores predichos por la serie se diferencien lo menos posible de los valores reales observados.
3) Que se obtengan residuales que constituyan un ruido blanco.

4) Que se usen tan pocos parámetros como sea necesario.

El primer objetivo se logra mediante la prueba *t* de Student para cada parámetro, y determina hasta que punto el parámetro en cuestión es significativamente diferente de cero. Esta prueba aparece en uno de los reportes del paquete estadístico SPSS.

El segundo objetivo se alcanza mediante el cálculo de algunos estadísticos de diagnóstico como son el error medio, los errores en porcientos y la suma de cuadrados de los errores, los cuales se pueden obtener utilizando un comando específico del propio SPSS

El cuarto objetivo, conocido como criterio de parsimonia, es en cierto sentido cuestionable cuando es la computadora quien hace las estimaciones y los pronósticos, pero en general, usar el menor número de parámetros facilitará la claridad y verificación del modelo y el pronóstico. Este principio de economía de la lógica formal, mundialmente conocido como *Ockham's Razor* (o navaja de Ockham) es inherente a varias formas del razonamiento bayesiano, como han referido e incluso formalizado varios autores (Jeffrey 1992; Jeffrey 1992) y establece en esencia, que entre dos modelos que produzcan resultados similares, no hay por qué acogerse al más complejo. En el caso de series temporales su aplicación se puede lograr con algunos estadígrafos que se comentan posteriormente (índices de Akaike o de Schwartz) y que permiten comparar el costo-beneficio de varios modelos y decidir hasta que punto "vale la pena" complicar un modelo con más parámetros en términos de los beneficios que de ello se obtienen.

Lograda en la práctica la estimación de los parámetros de una muestra, hay que validar hasta que punto el modelo estimado ajusta bastante bien la realización. Esto se conoce como fase de validación.

La parte más importante del diagnóstico es el chequeo de que los residuales constituyen realmente un ruido blanco. Ello significa que se debe probar estadísticamente que los residuales son no correlacionados, tienen media cero y varianza constante. En la práctica ello se logra con la graficación de los residuales como serie, el estudio de la función ACF(h) PACF(h) de esta, que debe mostrar en particular una estructura ARIMA (0,0,0)(0,0,0) y ciertos Q-estadísticos (conocidos como estadísticos Box-Ljung) que prueban la hipótesis nula de que esta serie no muestra autocorrelaciones significativamente diferentes de cero.

La diferencia aparente con el análisis de residuales de la regresión es que no se necesita probar que los residuales se distribuyen normalmente, ni siquiera que tengan la misma

distribución para cada instante de tiempo. Sin embargo la validez predictiva de los pronósticos depende en muchos casos de que los residuales tengan la misma distribución y la elaboración de los intervalos de confianza es más fácil si los residuales se distribuyen normalmente. Desde este punto de vista tiene interés saber adicionalmente si los residuales se ajustan a una distribución normal.

Existen otros criterios que ayudan a validar un modelo o incluso a seleccionar entre varios modelos el mejor. Por ejemplo, si el modelo tiene componentes de medias móviles es deseable que mantenga su inversibilidad pues de hecho, el pronóstico con estos modelos se hace "invirtiéndolos" hacia modelos autorregresivos. Otro criterio es el análisis de varianza que refleja la suma de cuadrados de los residuales y mientras menor, pues mejor es el modelo, y por último los criterios de Akaike o de Schwartz que permiten comparar un modelo obtenido con otros desde el punto de vista del ajuste y del costo en parámetros. Más concretamente:

- **AIC (Akaike information criterion)** es el estadístico de Akaike que ayuda a decidir si el orden del modelo es correcto. Este estadístico mide eficiencia (más bien ineficiencia, porque es **mejor mientras más pequeño**) en términos de ajuste-costo de los parámetros.
- **SBS (Schwartz bayesian criterion)** es un estadístico similar al de Akaike para decidir si el orden del modelo es correcto. El criterio de Akaike es más apropiado para modelos autorregresivos, mientras que el criterio bayesiano de Schwartz es preferible cuando los modelos que se comparan son medias móviles o mixtos.

También se considera el parámetro Log likelihood que es el logaritmo de la función de verosimilitud que mide la probabilidad de los resultados observados condicionada a los parámetros estimados. Más precisamente, es el logaritmo de la representación algebraica de la distribución de probabilidad conjunta de la muestra de datos en la cual los valores observados son tratados como fijos y los valores de los parámetros como variables. El logaritmo de la función de verosimilitud es una medida más fina de la verosimilitud (entre $-\infty$ y $+\infty$) y deseable lo mayor posible.

Otra fase que merece consideraciones teóricas importantes es la de pronóstico, la cual se analiza con rigor suficiente en (Fuller 1976). Ideas esenciales son las siguientes:

Dadas "n" observaciones de una realización se pretende predecir la observación "n+s" donde s es un entero positivo. A causa de la naturaleza funcional de una realización, la

predicción o pronóstico no es otra cosa que una extrapolación. Recuérdese que en el análisis de regresión clásico, las extrapolaciones son muy peligrosas y el mérito fundamental de la teoría de series de tiempo desde el punto de vista práctico es la posibilidad de brindar pronósticos más certeros fuera de los intervalos de valores observados, hacia delante o hacia atrás.

1.8. Metodología Box-Jenkins.

La Metodología de Box-Jenkins es realmente un proceso multipaso e iterativo de análisis de series de tiempo y pronósticos, y consiste esencialmente de cuatro fases:

- Identificación.
- Estimación de parámetros.
- Chequeo diagnóstico.
- Pronóstico.

La suposición de partida es que la serie de tiempo bajo análisis pertenece a una clase de modelos ARIMA. La metodología ayuda no solo a identificar un modelo, sino a perfeccionarlo en varias de sus fases. Es importante enfatizar que para un juego de datos específicos puede existir más de un modelo ARIMA que ajuste bien los datos. El chequeo diagnóstico incluye no solo el estudio de la validez de un modelo, sino también la comparación de varios posibles. El propio rastreo del pronóstico con nuevos valores disponibles hace que una serie pueda también mejorarse sucesivamente. Por razones de este tipo es que el proceso se define como iterativo aunque se distinguen las cuatro fases anteriormente mencionadas. Esta metodología no es un "algoritmo" pues no se garantiza siempre la convergencia a una solución. De hecho si la serie no es ARIMA o transformable a una tal serie, la metodología de Box-Jenkins no puede dar resultados.

Un "diagrama de flujo" que muestra las fases y el carácter iterativo se muestra en el **anexo I-5**.

Una vez que la serie es estacionaria, los correlogramas trazados permiten la identificación inicial del modelo. El proceso de identificación puede ser concebido como un proceso cíclico de aproximaciones sucesivas en el que intervienen en el primer nivel, las fases de identificación, estimación, y análisis de las autocorrelaciones de los residuales. En el segundo nivel, el ciclo abarca además el pronóstico y su contraste con valores reales.

La idea práctica de este primer lazo (ver diagrama) se puede fundamentar fácilmente con el llamado **principio para la identificación sucesiva** y se plantea que:

Si interesa escribir una serie X_t como un modelo ARIMA (p,d,q) y el residual E_t no resulta un ruido blanco, sino que E_t ε ARIMA(p',d',q'), entonces X_t ε ARIMA(p+p',d+d',q+q').

Gracias a este principio, se puede comenzar suponiendo el modelo con la estructura más simple entre las plausibles y analizando la correlación de los residuales determinar la posible necesidad de elevar el orden del modelo.

En la fase de identificación del modelo, el ploteo de la serie y los residuales permiten descubrir e identificar tendencias a la periodicidad, además de tendencias lineales, polinómicas, o violaciones del carácter estacionario de la serie por falta de homogeneidad de varianzas. Las transformaciones y/o diferenciaciones, permiten muchas veces lograr el carácter estacionario y esto explica el primer lazo. Los pasos sucesivos de la metodología parten de que se ha alcanzado estacionaridad de la serie.

La estimación de parámetros es la fase de construcción del modelo donde se calculan los valores específicos para cada uno de los parámetros. Ya que la serie de tiempo que se esta modelando es solamente una muestra o realización del proceso que ella representa, realmente lo que se calcula son estimativas muestrales de los verdaderos parámetros.

El diagnóstico comienza prácticamente con los estadísticos que surgen en la fase de estimación, tiene su centro en el estudio de la correlación de los residuales y se extiende hasta la etapa de pronóstico en el sentido siguiente.

Una práctica general y bastante usual al comenzar el estudio de modelos de series de tiempo, es reservar desde el principio una parte de los datos (por ejemplo, la última cuarta o quinta parte), para validar el modelo y emprender todo el análisis (identificación, estimación y diagnóstico) con la primera parte de los datos. El pronóstico sobre el periodo de validación y su comparación con los valores reales proporciona un criterio efectivo de cuan válidos son los pronósticos a partir del modelo estimado. En el período de validación pueden comprobarse tantos los pronósticos a corto plazo como los pronósticos a largo plazo. En el primer instante a pronosticar, el valor predicho se estima a partir de un cierto número de valores anteriores a él. A partir del segundo instante, se tienen dos alternativas: utilizar el valor real (que esté disponible) en el instante anterior, o utilizar el valor recién pronosticado para ese punto. En general, si se pronostica utilizando valores reales de la serie anteriores al instante actual, aunque estén dentro del periodo de validación, los

pronósticos serán más exactos y se valida con ello el pronóstico a corto plazo. Si para predecir el valor en un instante dado se utiliza sólo los valores reales que sirvieron de base en la estimación de la serie y los valores pronosticados de instantes anteriores al actual, se valida el pronóstico a largo plazo. En este último caso, se obtiene un pronóstico mucho más grosero porque la información real se agotará al cabo de ciertos pasos.

En el pronóstico real a largo plazo, esto es, sobre un período para el cual no se tienen valores reservados, es imprescindible utilizar después del primer paso, la información previamente pronosticada.

1.8.1. Metodología Box-Jenkins para series estacionales.

El análisis de una serie de tipo estacional ARIMA es una extensión del principio de identificación sucesiva. Salvo un detalle, que se explica inmediatamente, se trata primero de identificar y ajustar los parámetros como si fuera una serie estacional pura (P,D,Q)S y luego, estudiando los residuales, identificar y estimar los parámetros de la posible componente regular (p,d,q). El modelo definitivo será (p,d,q)(P,D,Q)S.

Existe otra vía alterna de trabajar una serie estacional (de una manera dual), pero se elige esta por las siguientes tres razones:

1. La dependencia estacional es determinante, más gruesa y requiere usualmente de menores valore de P,D,Q. Para su identificación y estimación más fina posible, es mejor trabajar con la serie original.
2. La identificación de un modelo ARIMA, parte siempre de la estacionalidad de la serie, lograda con transformaciones o diferenciaciones.
3. La estacionaridad alcanzada por transformaciones y diferenciaciones tanto regulares como estacionarias, permite estimar más claramente la constante μ como media de la serie estacionaria.

Así, el orden de identificación usual es realmente:

d-diferenciación.

S-estacionalidad de la serie.

D-diferenciación estacional.

Todo ello precedido posiblemente de transformaciones para alcanzar homocedasticidad (homogeneidad de varianzas) y con el objetivo final de alcanzar estacionaridad (en este

momento se podría estimar ya la constante μ). Una vez alcanzada la estacionaridad, se identifican sucesivamente:

(P,Q)-ordenes autorregresivo y de medias móviles estacional a partir de la serie transformada y diferenciada. La identificación de P y Q permite estimar los parámetros Φ_i, i=1,2,…,P y Θ_j, j=1,2,…,Q y calcular los residuales E_t de un modelo estacional supuestamente puro.

(p,q)-ordenes autorregresivo y de medias móviles regulares a partir de los residuales del procesamiento anterior. La identificación de "p" y "q" puede considerarse un afinamiento del modelo y permite estimar los parámetros φ_i, i=1,2,…,p y θ_j j=1,2,…,q y calcular los residuales e_t que es de esperar sea un ruido blanco.

De esta manera se mantiene válido el orden del flujo en el diagrama o metodología de Box-Jenkins.

Las funciones de autocorrelación estacional SACF(h) y la autocorrelación parcial estacional SPACF(h) existen y tienen las mismas apariencias para los diferentes valores de P y Q que en el caso regular. En particular:

- En una serie autorregresiva estacional de orden P, la función SACF(h) mostrará una rápida declinación a cero y la función SPACF(h) mostrará "P" espigas.
- En una serie de media móvil estacional de orden Q, la función SACF(h) mostrará "Q" espigas y la función SPACF(h) mostrará una rápida declinación a cero.
- En una serie mixta estacional, de orden (P,Q), que sea estacionaria, los patrones serán más complejos; pero ambas funciones mostrarán una rápida declinación a cero.
- Si en una serie estacional, la SACF (h) no muestra una rápida declinación a cero, ella no es estacionaria estacionalmente y probablemente requiera de una o dos diferenciaciones estacionales.

1.9. Análisis de intervención con modelos ARIMA.

El comportamiento histórico de un proceso se ve afectado frecuentemente por la influencia de un factor externo en un instante de tiempo dado o a patir de cierto momento, o en el intervalo comprendido entre dos ciertos instantes de tiempo.

Si tales procesos son modelables ARIMA, la serie correspondiente debe mostrar un "salto" o "cambio brusco" producto de esta intervención y es deseable entonces

"cuantificar" este salto, incluyéndolo en el modelo para que este responda mejor a la realización, y en particular determinar hasta que punto es significativo.

En matemática se utilizan frecuentemente las dos funciones siguientes para representar un salto discreto:

La función "paso"o "salto unitario" definida por:

$$u(t)= \begin{cases} 0, & si\ t < 0 \\ 1, & si\ t \geq 0 \end{cases}$$

La función "delta"o pulso unitario definida por:

$$\delta(t) = \begin{cases} 0, & si\ t \neq 0 \\ 1, & si\ t = 0 \end{cases}$$

Box y Tiao (Box 1975) generalizaron las series de tiempo ARMA objeto de estudio, a series de la forma:

$$(1-\varphi_1 B-...-\varphi_p B^p)X_t = (1-\theta_1 B-\theta_2 B^2-....-\theta_q B^q)e_t + \sum_{i=1}^{n} a_i \bullet r_i(t)$$

donde $r_i = r_i(t)$ son ciertas variables "regresoras" que cumplen determinadas condiciones y en particular que pueden ser del tipo: $r_i = u(t-t_i)$ o $r_i = \delta(\mathrm{t}-\mathrm{t_i})$ para ciertos t_i conocidos (i=1,2,…n). Los coeficientes "a_i" son nuevos parámetros a estimar. Nótese de hecho que un término constante en la serie, representando la media puede considerarse el coeficiente a_0 en el caso particular de $r_0 = u(t)$, (esto es, $r_0 = 1, t \geq 0$).

La forma en que se introducen los regresores en la estimación del modelo es bien importante y requiere de algunas consideraciones especiales, que se precisan en el siguiente epígrafe y se espera que sean útiles en otras situaciones más generales.

1.9.1. Sobre la introducción de regresores en modelos con diferenciación

Cuando tratamos con un modelo ARMA (p,0,q)(P,0,Q)S la introducción de los regresores en el SPSS puede lograrse fácilmente como variables independientes adicionales para lograr los resultados propuestos por Box y Tiao. Pero si hay alguna diferenciación regular y/o estacional, ellos también serán diferenciados y por tanto, el regresor que actúa sobre la

variable dependiente no es el introducido como variable independiente sino que es su diferencial discreta y por tanto sus efectos pueden ser muy diferentes de los esperados.

Los autores del SPSS proponen que en estos casos se calcule previamente las series diferenciadas de la variable dependiente y sea a estas series diferenciadas las que se le añadan los regresores, para evitar su diferenciación. Teóricamente esto está claro; pero desde el punto de vista práctico ello genera un problema posterior con el pronóstico por acumulación de errores. Si por ejemplo, la serie original X_t necesita ser diferenciada regular y estacionalmente, se tendría:

$$Y_t = X_t - X_{t-S} \qquad \text{y} \qquad Z_t = Y_t - Y_{t-1}$$

Cuando se busca el modelo para Z_t se tiene:

$$Z_t = \tilde{Z}_t + e_t$$

y el error e_t se arrastra y acumula en la integración hacia las series originales:

$$\tilde{Y}_t = \tilde{Z}_t + \tilde{Y}_{t-1} \qquad \text{con} \quad \tilde{Y}_1 = Y_1$$

y $$\tilde{X}_t = \tilde{Y}_t + \tilde{X}_{t-S} \quad \text{con} \quad \tilde{X}_i = X_i \quad \text{para} \quad i = 1,2,...,S$$

El problema se agrava si en el modelo de Z_t intervienen términos de medias móviles.

Este problema fue constatado concretamente en las series que se abordan en el presente trabajo y también en el Trabajo de Diploma (Monteagudo 2007), desarrollado paralelamte a este. La alternativa de solución es introducir como variables independientes las primitivas de los regresores que finalmente se desean, para que ellas sean diferenciadas y el pronóstico se haga directamente de la serie original. Así por ejemplo, si tenemos una serie que va a ser una vez diferenciada regularmente, y deseamos tener en un instante determinado t_o una función pulso: $\delta(t\text{-}t_o)$ entonces debemos introducir como variable independiente una función paso unitario $u(t\text{-}t_o)$ porque su derivada discreta es la función pulso deseada. Otras situaciones pueden ser más complejas, pero también solubles, como se ilustra en los ejemplos siguientes en los cuales consideramos, que tenemos datos trimestrales (estacionalidad S=4). Las construcciones se pueden generalizar a cualquier estacionalidad.

Primitiva regular y estacional de una función pulso

Se desea en este ejemplo construir una función f(t) tal que después de ser diferenciada regularmente una vez y diferenciada estacionalmente una vez, conduzca a la función $\delta(t)$. Los resultados podrán fácilmente ser trasladados después a cualquier punto t_o. Será

suficiente obtener $f(t)$ de manera que la serie diferenciada estacionalmente sea la función paso unitario $u(t)$, esto es:

$$f(t) - f(t-4) = u(t) \quad \forall t \in \Re$$

Observe que trabajamos con valores positivos y negativos de t para poder luego desplazar el centro a un punto t_o>0. El valor de f en el centro (en este caso t_o=0) puede definirse arbitrariamente pues de hecho la primitiva deseada está definida salvo una constante. Fijemos por ejemplo $f(0)$=1/4, esto es el inverso de la estacionalidad. La idea de definir $f(t)$ para t>0 es lograr que al cabo de 4 pasos se obtenga una diferencia de 1, Por tanto $f(1)$=1/2, $f(2)$=3/4, $f(3)$=1, $f(4)$=1+1/4, $f(5)$=1+1/2,…y en general f(t)=(t+1)/4 para todos los t mayores o iguales a 0. Así garantizamos que para valores mayores o iguales a 4 se tenga $f(t)$-$f(t$-4)=1=$u(t)$. Esta misma fórmula tiene que cumplirse para los valores de t=0,1,2,3. Por tanto f(t)=(t+1)/4 para los t mayores o iguales que -4. En particular f(-1)=0, f(-2)=-1/4, f(-3)=-1/2, f(-4)=-3/4. A partir de aquí, moviéndonos a la izquierda del eje, debemos tener f(-5)=f(-1)=0, f(-6)=f(-2)=-1/4, f(-7)=f(-3)=-1/2, f(-8)=f(-4)=-3/4 porque para los t negativos $u(t)$=0. Entonces se repiten los valores en ciclos de 4 valores:

f(-9) = f(-5)=0, f(-10)=f(-6)=-1/4, f(-11)=f(-7)=-1/2, f(-12)= f(-8) =-3/4,

f(-13)=f(-9)=0, f(-14)= f(-10)=-1/4, f(-15)= f(-11)=-1/2, f(-16)=f(-12)=-3/4,

etc. Está claro que estos 4 valores se determinan fácilmente por el número del trimestre que preceden al centro. La función así obtenida se grafica en el **Anexo I-6**.

Primitiva regular y estacional de una función paso unitario

Supongamos ahora que se desea construir una función g(t) tal que después de ser diferenciada regularmente una vez y diferenciada estacionalmente una vez, conduzca a la función $u(t)$. Será suficiente obtener $g(t)$ de manera que la serie diferenciada regularmente sea la función $f(t)$, encontrada en el epígrafe anterior, esto es:

$$g(t) - g(t-1) = f(t) \quad \forall t \in \Re$$

Fijamos arbitrariamente el valor en el centro, por ejemplo $g(0)$=$f(0)$=1/4. Entonces la relación anterior entre g y f permite calcular fácilmente los valores de $g(t)$. En efecto para los t>0, tendremos:

$$g(1) = g(0) + f(1) = \frac{1}{4} + \frac{1+1}{4}$$

$$g(2) = g(1) + f(2) = \frac{1}{4} + \frac{1+1}{4} + \frac{2+1}{4}$$

$$g(3) = g(2) + f(3) = \frac{1}{4} + \frac{1+1}{4} + \frac{2+1}{4} + \frac{3+1}{4}$$

$$g(4) = g(3) + f(4) = \frac{1}{4} + \frac{1+1}{4} + \frac{2+1}{4} + \frac{3+1}{4} + \frac{4+1}{4}$$

Y en general para los *t* mayores o iguales a cero, se tiene:

$$g(t) = g(t-1) + f(t) = \frac{1}{4} + \frac{1+1}{4} + \frac{2+1}{4} + \frac{3+1}{4} + ... + \frac{t+1}{4}$$

$$= \frac{t+1}{4} + \frac{1+2+3+...+t}{4} = \frac{t+1}{4} + \frac{t(t+1)}{8} = \frac{t^2+3t+2}{8}$$

La relación diferencial entre *g* y *f* permite obtener también los valores a la izquierda del centro. Así:

$$g(0) - g(-1) = f(0) = \frac{1}{4} \quad \Rightarrow \quad g(-1) = g(0) - \frac{1}{4} = 0$$

$$g(-1) - g(-2) = f(-1) = 0 \quad \Rightarrow \quad g(-2) = g(-1) = 0$$

$$g(-2) - g(-3) = f(-2) = -\frac{1}{4} \quad \Rightarrow \quad g(-3) = g(-2) + \frac{1}{4} = \frac{1}{4}$$

$$g(-3) - g(-4) = f(-3) = -\frac{1}{2} \quad \Rightarrow \quad g(-4) = g(-3) + \frac{1}{2} = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

$$g(-4) - g(-5) = f(-4) = -\frac{3}{4} \quad \Rightarrow \quad g(-5) = g(-4) + \frac{3}{4} = \frac{3}{4} + \frac{3}{4} = \frac{6}{4}$$

$$g(-5) - g(-6) = f(-5) = 0 \quad \Rightarrow \quad g(-6) = g(-5) = \frac{6}{4}$$

$$g(-6) - g(-7) = f(-5) = -\frac{1}{4} \quad \Rightarrow \quad g(-7) = g(-6) + \frac{1}{4} = \frac{6}{4} + \frac{1}{4} = \frac{7}{4}$$

etc. Podría buscarse incluso, una fórmula general utilizando ecuaciones en diferencias finitas; pero no es necesario pues es fácilmente observable que moviéndonos a la izquierda de -1, cada valor se obtiene del anterior, sumando en ciclo 0, ¼, ½, ¾, 0, ¼, ½, ¾,… y esto es fácilmente calculable en el SPSS utilizando el valor del trimestre. El gráfico de una tal función aparece en el **Anexo I-7.**

Primitivas de otros regresores necesarios.

En el estudio de la influencia del Período Especial sobre las series de consumo, surgió, además de los dos ejemplos anteriores, la necesidad de una variable independiente, llamémosla *h*(*t*), que diferenciada regularmente, produjera un valor 1 en un intervalo de

tiempo compacto [t_0,t_1] y el valor 0 fuera de este intervalo, esto es, una variable independiente cuya diferenciación regular $h(t)$- $h(t\text{-}1)$fuera igual al regresor $u(t\text{-}t_0)$- $u(t\text{-}t_1)$. Queda claro, de la metodología seguida anteriormente, que una vez fijado un valor, por ejemplo, $h(t_0)=1$, quedan automáticamente determinados los valores a la derecha de éste hasta $t=t_1$:

$$h(t_0+1)=2,$$

$$h(t_0+2)=3,$$

$$h(t_0+3)=4$$

Para cualquier punto intermedio entre t_0 y t_1 resulta

$$h(t_0+i)=i+1,...,$$

Y para t_1 se alcanza finalmente este "ascenso"

$$h(t_1)=h(t_0+(t_1-t_0))=(t_1-t_0)+1$$

Ahora, para valores $t> t_1$ fijamos $h(t)=(t_1\text{-}t_0)+1$, ello garantiza diferencias nulas a partir de $t= t_1+1$, mientras que para $t<t_0$ hacemos $h(t)=0$ para garantizar la diferencia unitaria en $t=t_0$ y nula a la izquierda de $t=t_0$.

En general, es muy simple obtener la primitiva de un regresor cuando aquella va a ser sometida a una (o varias) diferenciaciones regulares, como en el ejemplo anterior. Cuando aparecen diferenciaciones estacionales, o ambas, el razonamiento se complica ligeramente, pero en esencia se sigue la misma metodología descrita con anterioridad.

1.10. Tratamiento de outliers en análisis ARIMA. El comando RMV.

El análisis ARIMA para series que representen una o más observaciones extremas (outliers) es otro aspecto que se presenta en estas series. Como ocurre con la regresión en general, una observación de este tipo puede ser errónea o verdadera. Si es lo primero, se tienen las alternativas siguientes:

1) Considerarla un valor perdido y tratar de obtener la serie con este tipo de valor.
2) Intentar sustituir el outlier (como un valor perdido) por un valor probable determinado por cierta "interpolación" de valores de su entorno.

Si el valor perdido estuviera al final de la serie, no fuera tan grave "recortar" la serie hasta ese momento; pero si es una observación intermedia, esto no es factible, y en el análisis ARIMA se dificulta por ejemplo, el cálculo de los coeficientes de autocorrelación. La

sustitución de un valor perdido (o de un outlier previamente sustituido por un valor perdido) parece la técnica más fácil pero no es la única, y pocas veces hay implementaciones consecuentes en los paquetes de análisis de series de tiempo.

- Para sustituir el o los valores perdidos por valores más plausibles, el módulo TRENDS del SPSS dispone de un comando especial denominado RMV (viene de Remove Missing Values). Este comando define una nueva serie a partir de una existente con valores perdidos por el usuario o por el sistema y en la cual estos valores se han sustituido mediante algún criterio especificado.

Los modelos que se tratarán en los próximos capítulos, ayudaran a la comprensión de la metodología anteriormente expuesta, incluyendo el uso del regresor y el tratamiento de los outliers.

1.11. Como comparar las series utilizando el comando MANOVA y los datos matriciales.

Dados 2 o más municipios con series que respondan a modelos idénticos ARIMA(p,d,q)(Sp,Sd,Sq)S, similares por la estacionalidad, la diferenciaciones y los parámetros que se determinan, tenemos interés en comparar los parámetros desde un punto de vista multivariado para determinar hasta que punto dichos modelos se diferencian o no significativamente (los parámetros pueden incluir además posibles regresores de intervención). Está claro que esta comparación no puede ser realizada de forma univariada, parámetro a parámetro porque entre ellos hay correlaciones. El análisis multivariado es imprescindible.

Por razones prácticas, tenemos interés en realizar dicho Análisis Multivariado con el procedimiento MANOVA del SPSS, porque este nos brinda los posibles datos iniciales para hacerlo; pero además nos esclarece como hacerlo en cualquier otro paquete. En principio, se puede hacer un análisis multivariado a partir solamente de un conjunto de datos que abarcan los volúmenes de las muestras, los valores medios en cada grupo, sus desviaciones estándar, y sus matrices de correlaciones, adecuadamente estimadas. ¿Pero como ellos se estiman y se integran en un paquete como el SPSS?.

Desde muchas versiones anteriores a SPSS 9.0, los procedimientos de Análisis de Varianza (en particular MANOVA) permitían la entrada de datos en forma matricial para hacer dicho análisis en términos de un número mínimo de datos concentrados, en sustitución del fichero

clásico original. Gracias a ello, ya este tipo de análisis comparativo de series se había logrado en la UCLV con aquellas versiones anteriores del SPSS, por ejemplo en (Tarrau Brito 1996) se hizo la comparación de series de tasas provinciales de algunas enfermedades de declaración obligatoria entre diferentes grupos de edades siguiendo esta vía. Se utilizó allí un procedimiento original implementado en el SPSS sobre las base de sus posibilidades de trabajo con datos matriciales.

Actualmente el MANOVA del SPSS en versiones posteriores trabaja en forma ligeramente diferente con datos matriciales y el procedimiento seguido en (Tarrau Brito 1996) no es aplicable. En primer lugar la preparación de los datos matriciales dista bastante en su forma en las nuevas versiones del SPSS respecto a las anteriores, de hecho, los datos se preparan como un fichero .SAV, no como un fichero texto y con ciertas variables-sistema y formatos previamente definidos. En segundo lugar, el contenido del fichero es diferente. La cuestión esencial es que se trabaja no solamente con las desviaciones (varianzas) conjuntas entre los grupos, sino también con coeficientes de correlación "conjuntos". No resultaba claro como ellos se estimaban en el SPSS a partir de las correlaciones internas en cada grupo. En otras palabras, aunque teóricamente existen diferentes técnicas de estimación de estos valores conjuntos, no se dispone de la información de a cuál técnica se ajusta la estrategia del procedimiento MANOVA del SPSS.

Se sabe en (Tarrau Brito 1996) que en un MANOVA del SPSS la estimación de las varianzas en cada grupo de volumen n se utiliza la "varianza corregida", esto es un estimador no sesgado (dividiendo por n-1) y no de máxima verosimilitud (que se obtendría dividiendo por n). Consecuentemente la varianza conjunta debería ser calculada promediando las varianzas corregidas de los grupos, ponderada con volúmenes de las muestras de cada grupo menos 1. De cualquier manera deberíamos cerciorarnos de ello.

Pero además, eso hizo suponer que la correlación conjunta, vista como cociente de una covarianza por desviaciones estándar de las variables, exigiría en algún sentido, estimar la covarianza conjunta en forma no sesgada (en las que intervinieran los volúmenes de cada grupo menos 1) y las desviaciones estándar corregidas. Pero no es obvio, pues otras estimaciones hubieran sido posibles.

Experimentado varias alternativas fue posible encontrar y demostrar como se realizan estas estimaciones en el procedimiento MANOVA del SPSS, cuyas bases coincidieron con las suposiciones anteriores. Entonces, fue posible establecer fórmulas que permitieran, a partir

de estadísticos descriptivos a nivel de grupos, disponibles con cualquier paquete de software, estimar los estadísticos conjuntos necesarios para obtener el Análisis Multivariado con absoluta fiabilidad, en el SPSS, o en cualquier otro paquete estadístico.

1.11.1 Procedimiento general.

1ro. Se debe preparar un fichero de datos (fichero de datos .SAV estándar del SPSS) con los datos matriciales.

Este fichero tiene obligatoriamente las siguientes variables:

Rowtype_. Variable de cadena corta (variable sistema del SPSS con ese nombre exacto) que puede tomar los valores: **N**, **Mean**, **StdDev** o **Corr** e identifica el tipo de datos de las filas, que son vectores.

varGroup. Variable nominal que identifica los grupos que se van a comparar, puede tener el nombre que usted desea y puede ser etiquetada (por ejemplo la variable puede llamarse Municipio, y tener los valores 1: Caibarién 2: Placetas). Ella tomará valores solamente en las filas que se refieren a cada uno de los grupos, no a las filas que se refieren a datos globales

Varname_. Variable de cadena corta (variable sistema del SPSS con ese nombre exacto) y que en cada fila de la matriz de correlaciones tomará como valores uno de los nombres de las variables dependientes que usted va a comparar por grupos para especificar respecto a que variable trata esa fila.

Vardep1, Vardep2, Vardep3,…etc. Son los nombres de las variables dependientes (en forma de cadenas cortas) que usted va a comparar. En el caso de la comparación de series serían los nombres de los parámetros, por ejemplo **AR1, MA1, SMA1**. Estos son a su vez los valores posibles de la variable **Varname_.**

Los datos del fichero tendrían el aspecto siguiente (suponemos 3 variables dependientes y 2 grupos):

Rowtype_	VarGroup	Varname_	Vardep1	Vardep2	Vardep3
N			N_1	N_2	N_3
N	1		n_{11}	n_{21}	n_{31}
N	2		n_{12}	n_{22}	n_{32}
Mean	1		$mean_{11}$	$mean_{21}$	$mean_{31}$
Mean	2		$mean_{12}$	$mean_{22}$	$mean_{32}$
StdDev			Sp_1	Sp_2	Sp_3
Corr		Vardep1	1	$corr_{12}$	$corr_{13}$
Corr		Vardep2	$corr_{12}$	1	$corr_{23}$
Corr		Vardep3	$corr_{13}$	$corr_{23}$	1

En general, el fichero tiene los siguientes datos:

- Un vector **N** con los totales de casos de cada una de las variables, para calcular el coeficiente de correlación (es el total de casos válidos de cada variable)
- Un vector **N** de *n*'s para cada grupo que abarca los valores válidos de cada variable en ese grupo. La suma por columna de dichos vectores debe conducir a la primera fila (en el ejemplo $\mathbf{N_1=n_{11}+n_{12}}$, $\mathbf{N_2=n_{21}+n_{22}}$, $\mathbf{N_3=n_{31}+n_{32}}$)
- Un vector **Mean** de *mean*'s para cada grupo que abarca los valores medios de cada variable en ese grupo.
- Un vector **StdDev** de *Sp*'s con las *desviaciones estándar conjuntas* de cada variable para la muestra en general (conjunta entre los grupos). Ver después en detalle como se calculan las componentes de este vector a partir de las desviaciones estándar de las variables por grupos.
- Una matriz con vectores **Corr** de *correlaciones conjuntas* entre las variables dependientes. En cada fila se utiliza el nombre de la variable dependiente Esta matriz es simétrica y tiene 1's en la diagonal. Ver a continuación detalles de cómo se calcula la matriz de correlaciones conjuntas a partir de la matriz de correlaciones en cada grupo.

Sobre las desviaciones estándar conjuntas:

Normalmente, para cada variable dependiente se tiene el valor de la desviación estándar en cada grupo por separado y se requiere aquí la desviación estándar conjunta (en todos los grupos). Sea X una variable dependiente fijada, y supongamos que tenemos k grupos con tamaños no necesariamente iguales: $n_1, n_2, \ldots, n_k$ y en cada grupo tenemos la desviación estándar de X, de manera que tenemos $S_{X1}, S_{X2}, \ldots, S_{Xk}$, o equivalentemente, tenemos las varianzas $S_{X1}^2, S_{X2}^2, \ldots, S_{Xk}^2$. Entonces la varianza conjunta de X , que denotamos S_{Xp}^2 (usamos p para recordar pooled-variance) se calcula según un promedio de las varianzas ponderado por el volumen de las grupos (realmente los volúmenes menos 1, porque se supone que las desviaciones estándar de que se parte en cada grupo han sido así corregidas). Consecuentemente la desviación estándar conjunta es la raíz cuadrada de lo anterior, esto es.

$$S_{X_p} = \sqrt{\frac{\sum_{i=1}^{k} (n_i - 1) * S^2_{X_i}}{\sum_{i=1}^{k} n_i - k}}$$

Si en particular todos los grupos tuvieran el mismo tamaño, esto es $n_1 = n_2 = \ldots = n_k = n$, se reduce a la raíz cuadrada del promedio simple de las varianzas:

$$S_{X_p} = \sqrt{\frac{\sum_{i=1}^{k} S^2_{X_i}}{k}}$$

Así por ejemplo, si tuviéramos dos grupos, tendríamos en general

$$S_{X_p} = \sqrt{\frac{(n_1 - 1) * S^2_{X_1} + (n_2 - 1) * S^2_{X_2}}{(n_1 + n_2 - 2)}}$$

Y si $n_1 = n_2 = n$

$$S_{X_p} = \sqrt{\frac{S^2_{X_1} + S^2_{X_2}}{2}}$$

Sobre las correlaciones conjuntas:

Normalmente, para cada par de variables dependientes se tiene el valor de la correlación en cada grupo por separado y se requiere aquí la correlación conjunta (en todos los grupos). Sean X e Y dos variables dependientes fijadas, y supongamos que tenemos k grupos con tamaños no necesariamente iguales: n_1, n_2, …,n_k y en cada grupo tenemos la correlación de X y Y, de manera que tenemos $Corr_{XY1}$, $Corr_{XY2}$, …, $Corr_{XYk}$. Supongamos que tenemos las varianzas S_{X1}^2, S_{X2}^2, …, S_{Xk}^2 y también S_{Y1}^2, S_{Y2}^2, …, S_{Yk}^2 y por tanto, de acuerdo a lo anterior podemos tener las varianzas conjuntas de ambas variables: S_{Xp}^2 y S_{Yp}^2. Entonces la correlación conjunta de X e Y, que denotamos $Corr_{XYp}$ se calcula también según un promedio de las covarianzas ponderado por el volumen de los grupos dividido por las varianzas conjuntas: Más precisamente:

$$Corr_{XY_p} = \frac{\frac{\sum_{i=1}^{k} (n_i - 1) * Corr_{XY_i} S_{X_i} S_{Y_i}}{\sum_{i=1}^{k} n_i - k}}{\sqrt{\frac{\sum_{i=1}^{k} (n_i - 1) * S^2_{X_i}}{\sum_{i=1}^{k} n_i - k}} \sqrt{\frac{\sum_{i=1}^{k} (n_i - 1) * S^2_{Y_i}}{\sum_{i=1}^{k} n_i - k}}} = \frac{\sum_{i=1}^{k} (n_i - 1) * Corr_{XY_i} S_{X_i} S_{Y_i}}{(\sum_{i=1}^{k} n_i - k) S_{X_p} S_{Y_p}}$$

Si en particular todos los grupos tuvieran el mismo tamaño, esto es $n_1 = n_2 = \ldots = n_k = n$, se reduce al cociente del promedio simple de las covarianzas por las desviaciones conjuntas:

$$Corr_{XY_p} = \frac{\sum_{i=1}^{k} Corr_{XY_i} S_{X_i} S_{Y_i}}{k * S_{X_p} * S_{Y_p}}$$

Así por ejemplo, si tuviéramos dos grupos, tendríamos en general

$$Corr_{XY_p} = \frac{(n_1 - 1) * Corr_{XY_1} S_{X_1} S_{Y_1} + (n_2 - 1) * Corr_{XY_2} S_{X_2} S_{Y_2}}{(n_1 + n_2 - 2) * S_{X_p} * S_{Y_p}}$$

Y si $n_1 = n_2 = n$

$$Corr_{XY_p} = \frac{Corr_{XY_1} S_{X_1} S_{Y_1} + Corr_{XY_2} S_{X_2} S_{Y_2}}{2 * S_{X_p} * S_{Y_p}}$$

Un tal fichero puede crearse en el SPSS como fichero .SAV (esto es lo que necesitamos precisamente para la comparación de las series y para ello necesitaremos las fórmulas anteriores). Pero también podría obtenerse como salida adicional de un MANOVA a partir de un fichero de datos clásico y cargarse en cualquier momento con el comando **GET FILE** ….Aunque hay cierta flexibilidad en el orden de estas filas, se recomienda que la primera fila sea efectivamente el vector N de totales generales y que las filas Corr que componen la matriz de correlación sean consecutivas. No añadir filas de más, porque no serán utilizadas y hasta pueden confundir.

2do. Una vez que se tiene activo en el SPSS el fichero anteriormente descrito:

Se ejecuta el siguiente comando a nivel de sintaxis.

```
MANOVA
  Vardep1 Vardep2…  BY Vargroup(1,m)
[ / PRINT=CELLINFO(MEANS) ERROR ]
/ MATRIX=IN(*)
```

Esto resulta en el análisis de varianza multivariado y además un análisis de varianza univariado para cada variable dependiente. El subcomando PRINT es opcional: el reproduce las medias por grupos y la matriz de errores que tiene en el triángulo inferior las correlaciones y en la diagonal las desviaciones estándar. Sirve solo para comprobar que se están entendiendo bien los datos.

En el ejemplo podría ser

```
MANOVA
  Vardep1 Vardep2 Vardep3 BY Vargroup(1,2)
/ PRINT=CELLINFO(MEANS) ERROR
/ MATRIX=IN(*).
```

De esta manera se logra la comparación de las series deseadas en diferentes municipios y se puede abordar el problema de la comparación de series de diferente naturaleza detectadas en grupos independientes, en otras aplicaciones

Conclusiones parciales

En este capítulo se muestra, de manera resumida, los conceptos fundamentales de la modelación de series de tiempo. Se explican también las cuatro fases de la metodología Box-Jenkins. Además de la teoría establecida, se fundamentan procedimientos sobre como introducir óptimamente regresores de frecuente uso en la modelación de series que suponen diferenciación regular y/o estacional, así como procedimientos para hacer análisis multivariados que pretendan comparar series temporales obtenidas en grupos independientes. Estas fundamentaciones teóricas y recomendaciones prácticas, pueden tener aplicaciones en muchos otros campos, y por tanto, un alcance mucho más allá del presente trabajo.

Capítulo 2: Análisis de series según el enfoque clásico.

En este capítulo se trata brevemente el contenido referente a la teoría de series de tiempo según el enfoque clásico surgido en los años 20 del siglo pasado para luego lograr modelos matemáticos según este enfoque.

En una serie cronológica se presentan por lo general, varios tipos de movimientos o formas de comportamientos que son llamados componentes. Las componentes más frecuentes en la práctica son la tendencia, la estacionalidad, el ciclo y la perturbación aleatoria o error, los tres primeros con un carácter sistemático, y el último aleatorio. Posteriormente se ofrece una descripción intuitiva de los mismos que permita continuar con la presentación de los modelos para el análisis de las series de consumo de energía eléctrica de Santa Clara y Provincial respectivamente.

Las componentes de una serie cronológica son a su vez series cronológicas con características especiales que se combinan de diferentes formas para formar o componer la serie original. Estamos hablando entonces de un modelo del tipo:

$$\mathbf{Y_t = f(T_t, S_t, C_t, E_t)}$$

en el cual integramos las componentes de tendencia (T_t), estacionalidad (S_t), posibles ciclos (C_t) y errores (E_t)

2.1 Tendencia.

La tendencia representa el movimiento a más largo plazo de la media de la serie, el cual puede ser considerado como el valor en torno al cual se mueven los demás componentes. Es altamente característico de las series cronológicas de fenómenos sociales, en particular de series económicas. Se denota usualmente por T_t.

2.2. Estacionalidad.

Esta componente tiene gran importancia en aquellas series cronológicas cuya extensión exceda a varios años y que reflejan el comportamiento de objetos sobre los que tienen influencia, ya sea directa o indirecta, las diferentes épocas del año. La característica de la componente estacional radica en que esta es una componente periódica de período igual a un año o trimestre por lo general. Los factores que se reflejan en esta componente tienen un comportamiento más inestable que los que rigen en la tendencia, por eso resulta de tanta importancia un estudio correcto de esta componente. La estacionalidad es particularmente

importante en las series económicas y meteorológicas. La estacionalidad se representa por $S_{t.}$

2.3. Ciclo.

El ciclo es un movimiento oscilatorio u ondulatorio en la serie cronológica que se diferencia de la estacionalidad sólo en la periodicidad, pues mientras la estacionalidad tiene un período de un año, el ciclo no tiene una periodicidad fija y siempre tendrá una longitud de varios años. Esta componente recoge el efecto de una gran parte de factores estables e inestables que provocan este comportamiento de periodicidades variables.

La componente cíclica a pesar de hallarse presente en muchas de las series cronológicas, no siempre resulta de inclusión necesaria y consciente en los análisis pues para los pronósticos a corto plazo, no es imprescindible. La componente cíclica es extraordinariamente importante en series geográficas (meteorología y sismología). El ciclo se denota por $C_{t.}$ Por ejemplo, en Meteorología, el Ciclo puede estar representado por las influencias del "Niño", cada 5 años, o quizás de la "Niña.

2.4. Perturbaciones aleatorias.

Esta componente recoge en sí todo el comportamiento no sistemático de la serie, es decir, su movimiento irregular, el cual es aleatorio o debido a factores de alta influencia pero de naturaleza esporádica como pueden ser: guerras, terremotos, inundaciones, sequías, huracanes, etc. El carácter aleatorio de estos factores o de su aparición hace prácticamente imposible su pronóstico riguroso por lo que a la componente aleatoria se le da un tratamiento de variable aleatoria para poder estimar su valor central y calcular límites de confianza para el desenvolvimiento futuro de esta variable. Se denota esta componente por $E_{t.}$

Estas cuatro componentes que no son más que cuatro series cronológicas, se combinan en determinada forma para dar lugar a la serie original. A la forma en que se combinan estas series se le denomina "modelo", siendo de especial importancia los modelos aditivo y multiplicativo:

$\mathbf{Y_t = T_t + S_t + C_t + E_t}$ **Modelo aditivo.**

$\mathbf{Y_t = T_t \,.\, S_t \,.\, C_t \,.\, E_t}$ **Modelo multiplicativo.**

El problema central que se presenta en la investigación de una serie cronológica mediante este enfoque, es la identificación de las componentes de forma tal que puedan ser aisladas y estimadas por separado para luego hacer predicciones con uno u otro modelo.

2.5. Estimación de las componentes de una Serie Cronológica.

La estimación de estas componentes exige primeramente la adopción de un modelo y a partir de él, ir estimando componente a componente y eliminando su efecto cada vez que una es estimada. El orden de estimación más difundido es: primero estacionalidad, luego tendencia, seguidamente el ciclo, si se va a estudiar, y finalmente los residuos o perturbaciones aleatorias.

Con el fin de "suavizar" la serie original se usa como filtro, el método de los promedios móviles de orden s, donde s puede ser cualquier valor mayor o igual a uno. En tanto s es mayor, la suavización es más acentuada. En las series estacionales se usa frecuentemente como valor de s, la estacionalidad por ejemplo, el número de subperíodos del año, a saber, dos si es semestral, cuatro si es trimestral, doce si es mensual etc. Con este método se suaviza un poco la serie, atenuándose la componente aleatoria y las estacionales.

Profundizando lo anterior, la idea surge de crear a partir de la realización original:

$X_1, X_2, X_3 \ldots\ldots\ldots X_n$

una nueva realización:

$Y_1, Y_2, Y_3, \ldots\ldots\ldots\ldots Y_{n-1}$

en la que cada $Y_i = \dfrac{(X_i + X_{i+1})}{2}$

la cual recibe naturalmente el nombre de serie de "medias móviles" de la original. Sobre la base de condiciones bastantes generales, la nueva serie presenta "menos fluctuaciones" que la original, y hasta cierto punto caracteriza mejor su "tendencia". Más precisamente, una serie de medias móviles caracteriza "tendencias a largo plazo + ciclos en tendencia". Si se dispone de bastantes valores en la realización (n es grande), la serie de medias móviles puede ser a su vez "suavizada" calculando una tercera serie, media móvil de la segunda, y así sucesivamente (Cué Muñiz 1987).

2.5.1. Separación de la componente estacional.

La tendencia estacional se puede considerar como la distribución de la forma en que cada momento dentro de una "estación" o "período" modifica el nivel general de la serie.

La separación del "factor o componente estacional" permite no sólo precisar la tendencia estacional de la serie, sino precisar además la tendencia y los residuales de la serie, (libres de influencias estacionales). A la serie "purificada" del efecto estacional es lo que se llama "serie desestacionalizada".

El proceso de separación de la tendencia estacional: S_t, la tendencia regular desestacionalizada: T_t (tendencia propiamente dicha y pequeños ciclos) y los residuales E_t, requiere conocer la estacionalidad "s" de la serie y además, el modelo al que responde, esto es, la forma en que cada una de estas componentes interviene en el nivel de la serie general, que se denota por Y_t.

En procesos ARIMA estacionales se utilizan nociones análogas para separar componentes estacionales (en forma multiplicativa), luego componentes regulares y finalmente residuales, a partir del conocimiento de la estacionalidad y una previa identificación del modelo.

A continuación se ilustran ideas fundamentales de cómo separar estas componentes en el caso de una serie no necesariamente ARIMA. La teoría parece aritmética y simple. Ciertamente es mucho menos sofisticada que la correspondiente a modelos ARIMA estacionales; pero tiene una fundamentación no trivial basada en la teoría de estimación que aparece desarrollada teóricamente en (Akaike 1974) y (Schwartz 1976). Estos procedimientos fueron implementados después con el nombre de Census Method I y son utilizados en el SPSS para la descomposición de series en dos tipos de modelos básicos: los aditivos y los multiplicativos ya mencionados anteriormente.

Así, en el modelo aditivo si se nombra M_t a la serie de promedios móviles con amplitud de la estacionalidad, esta sólo tendrá componentes de tendencia y ciclo, o sea $M_t = T_t + C_t$ ya que el componente aleatorio y estacional se atenúa en el cálculo de los promedios móviles.

Si se resta de la serie original Y_t a la serie de promedios móviles M_t se obtiene una nueva serie que sólo presenta las componentes estacional S_t y aleatoria E_t, es decir:

$$Y_t - M_t = S_t + E_t$$

Si se tabula esta nueva serie por estaciones, por ejemplo, por trimestre (debe recordarse que con el método de los promedios móviles se pierden datos originales, (Cué Muñiz 1987) y se promedian estos, se obtienen los coeficientes de estacionalidad para cada estación. En el estudio dentro de este capítulo se verá cómo se realiza este proceso en el SPSS.

También en el modelo multiplicativo si se nombra M_t a la serie de promedios móviles con amplitud de estacionalidad para cada estación, ésta sólo tendrá componentes de tendencia y ciclo, pero aquí $M_t = T_t \cdot C_t$. Si se dividen todos los elementos de la serie original Y_t por los de la serie de promedios móviles M_t se obtiene una nueva serie que sólo presenta las componentes estacional S_t y aleatoria E_t es decir:

$$\frac{Y_t}{M_t} = S_t \cdot E_t$$

Si se tabula esta nueva serie por estaciones y se promedian éstos, se obtienen los coeficientes (índices) de estacionalidad para cada estación. Se verá cómo se realiza este proceso en el SPSS también en las series del presente estudio. En el modelo aditivo la suma de los coeficientes estacionales debe ser cero; en el modelo multiplicativo la suma de los índices estacionales debe ser igual a la estacionalidad (multiplicada por 100 si se expresa en %), lo cual equivale a que el promedio de ellos sea 100. En caso contrario la estimación se encarga de hacer la corrección.

En el modelo multiplicativo $\mathbf{Y_t = T_t \,.\, S_t \,.\, C_t \,.\, E_t}$, el efecto estacional es proporcional al nivel de la serie. Esto es, la influencia estacional en cada momento de un periodo típico, digamos un trimestre con estacionalidad 4, es característica del trimestre y actúa multiplicando la tendencia general de la serie, de manera que la variación estacional es más acentuada, en tanto los valores generales de la serie son mayores.

Cuando el efecto del factor estacional no depende del nivel general de la serie, sino que simplemente, se “suma”a la tendencia, el modelo se identifica como aditivo, o sea,

$\mathbf{Y_t = T_t + S_t + C_t + E_t}$.

La identificación de un modelo aditivo o multiplicativo, así, como la estacionalidad, deben hacerse a partir del ploteo de la serie y el análisis descriptivo general utilizando las ideas anteriores. A veces puede confundirse la necesidad de aplicar una transformación

logarítmica para ganar estacionalidad (hacer constante la varianza), con la necesidad de utilizar un modelo multiplicativo estacional. Es una confusión admisible. De hecho, la aplicación de una transformación logarítmica, transforma una multiplicación de efectos en una suma de ellos; pero debe comprenderse intuitivamente la diferencia entre los modelos.

a) $Y_t = T_t + S_t + Z_t$

b) $Y_t = T_t \cdot S_t \cdot Z_t$

c) $Ln(Y_t) = T_t + S_t + Z_t$

d) $Y_t = Ln(T_t) + Ln(S_t) + Ln(Z_t)$

Cuando se estudian series estacionarias, a veces el ploteo de la serie refleja una violación de esta hipótesis (varianza constante) porque muestra grandes oscilaciones cuando los valores de la serie eran más grandes en valor absoluto que cuando eran más pequeños. La transformación logarítmica frecuentemente permitía alcanzar la homocedasticidad deseada en tales casos.

Sin embargo, en el procedimiento que se utiliza en el presente trabajo, no se parte de que la serie Y_t es estacionaria. Por tanto, ninguna transformación de la serie exigida a priori, y aún cuando se haga, se esta hablando ya de Y_t como la serie transformada. Otra cosa es que esta responda a un modelo aditivo o multiplicativo respecto a sus componentes regulares, estacionales y residuales, de forma que se están considerando sólo modelos de la forma a) y b).

El modulo TRENDS del SPSS tiene una opción SEASON, que realiza la descomposición de un modelo aditivo o multiplicativo utilizando los métodos descritos brevemente en los párrafos anteriores.

2.5.2. Uso básico de la regresión en las series de tiempo.

En muchos casos es posible especificar la serie, o al menos, la media de la serie, como una función simple del tiempo, por ejemplo un polinomio de bajo orden, o una función trigonométrica de t, y en tales casos se utiliza la regresión clásica para lograr esta especificación.

El uso más simple de la regresión se hace en aquellos casos en que la serie misma responde a un modelo de regresión, más detalles consultar el epigrafe 2.5.2 de (Mora Villegas 2003)**.**

Para facilitar el uso de las regresiones en el tiempo, reducibles a regresiones lineales, el modulo TRENDS del SPPSS, brinda el comando de regresión CURVEFIT que permite estudiar rápidamente el ajuste de varios modelos a una serie dada.

La sintaxis básica de CURVEFIT supone especificar la variable que define la serie y el modelo entre un conjunto de ellos, entre los cuales se encuentran:

Modelo	Ecuación	Transformación de linelización
LINEAR	$X = b_0 + b_1 t$	$f_0(x) = x$, p=1 $f_1(t) = t$, $\beta_j = b_j$,j=0,1
LOGARITHMIC	$X = b_0 + b_1 \ln(t)$	$f_0(x) = x$, p=1 $f_1(t) = \ln(t)$, $\beta_j = b_j$,j=0,1
INVERSE	$X = b_0 + \frac{b_1}{t}$	$f_0(x) = x$, p=1 $f_1(t) = \frac{1}{t}$, $\beta_j = b_j$,j=0,1
CUBIC	$X = b_0 + b_1 t + b_2 t^2 + b_3 t^3$	$f_0(x) = x$, p=1 $f_1(t) = t$, $f_2(t) = t^2$, $f_3(t) = t^3$ $\beta_j = b_j$,j=0,1,2,3

CURVEFIT crea en el fichero activo series de pronósticos, de errores, de intervalos de confianza, etc, que pueden ser comparadas con las otras series producidas por otros comandos del modulo TRENDS, por ejemplo con FIT.

2.5.3. Estimación de la tendencia.

Ya visto el problema de estimar la estacionalidad S_t según se trató brevemente en 2.5.1, y de aislarla del resto de las componentes, debe abordarse la estimación de otras componentes; la componente que debe estimarse a continuación es la tendencia, la cual se estima a partir de la serie estacionalmente ajustada.

La regresión contra el tiempo puede ser utilizada además para caracterizar una componente de la serie, como su “media general” o “tendencia”. A continuación se precisa un poco este concepto.

La teoría ARIMA que se presentó en el capítulo precedente y mucho acerca de la teoría general de las series, se basa en la consideración de que la serie es estacionaria.

Muchas series en la práctica no lo son y ello puede suceder por cualquiera de las razones siguientes:

1) La media de la serie es una función del tiempo, diferente de una constante.
2) La varianza es una función del tiempo, diferente de una constante.

3) La serie es generada por un tercer tipo de mecanismo estocástico no estacionario, por ejemplo, la serie pudiera ser la suma de dos variables aleatorias independientes.

Analizando el modelo aditivo veremos que se puede expresar como:

$Y_t = T_t + S_t + E_t$.

Donde T_t es la componente de "tendencia", S_t es la componente "estacional" y E_t es la componente "irregular" o "aleatoria". En nuestra terminología, E_t es una serie de tiempo estacionaria y T_t es descompuesto en ocasiones a su vez en: $T_t = L_t + C_t$, donde L_t es una componente de tendencia esencial o a "largo plazo" y C_t la tendencia "cíclica" que explica pequeñas ondas o fluctuaciones de la tendencia, independientes de la componente estacional.

Ahora interesa mostrar como la regresión contra el tiempo puede ser utilizada para caracterizar la tendencia general de la serie.

Una primera idea, sería, utilizar como tal la serie e intentar una regresión respecto a ciertas funciones fijadas del tiempo, por ejemplo, polinómica, exponencial, etc. Aún conociendo, que dicha función no es capaz de explicar las fluctuaciones de la serie, el método de los mínimos cuadrados ordinarios daría la curva de este tipo "que más se aproxima". **Pero en esta variante hay un problema;** ello equivale a admitir que la serie de tiempo responde a un modelo de la forma: $\vec{Y} = \psi \vec{\beta} + \vec{E}$ en la cual el error $\vec{E}$ no es ya, *un ruido blanco distribuido normalmente*, sino una serie de tiempo, en el mejor de los casos estacionaria. El término $\psi \vec{\beta}$ identifica en este caso la tendencia a largo plazo, y por tanto $\vec{E}$ abarca los pequeños ciclos y componentes irregulares.

Lo anterior implica entonces, a la luz del presente trabajo, que cuando se logre un modelo, aditivo o multiplicativo, la serie E_t tiene que cumplir que la misma al menos sea una serie ARMA, de lo contrario el modelo no es válido. De ahí entonces, que un estudio serio de los modelos clásicos requiera, aún cuando históricamente surgió después, de la teoría ARIMA para series de tiempo.

En resumen, se tienen dos métodos para separar la componente de tendencia de la serie y escribirla como función de tiempo con ayuda de la regresión:

1) Aplicar regresión de la serie original con algunas de las funciones contempladas en CURVEFIT.

2) Obtener una realización de la tendencia suavizando la serie original con medias móviles y aplicar regresión a esta realización para expresarla como una función conocida del tiempo, **consultar el epígrafe 2.5.3 de (Mora Villegas 2003).**

Como ejemplo de implementación del estudio brevemente explicado anteriormente, se tomará en primer lugar, la serie de energía eléctrica del municipio de Santa Clara**.**

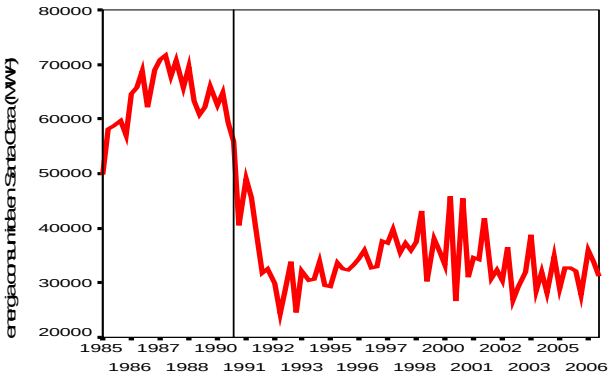
2.6. Serie energía eléctrica consumida por Santa Clara.

Con esta serie se pretende:

1) Calcular los coeficientes estacionales e índices estacionales.
2) Determinar la tendencia de la serie desestacionalizada por el modelo aditivo y el modelo multiplicativo.
3) Estimar los modelos aditivo y multiplicativo.

2.6.1. Detección de outliers y desestacionalización de la serie de Santa Clara.

Primeramente se entran los datos de la serie original con el nombre Santa..e y cuyo gráfico es:



Luego se define la fecha la cual se activa con las opciones **Datos \ definir fechas \Años, trimestres.**

```
The following new variables are being created:
  Name         Label
  YEAR        YEAR, not periodic
  QUARTER      QUARTER, period 4
  DATE_        DATE.  FORMAT:  "QQ YYYY"
```

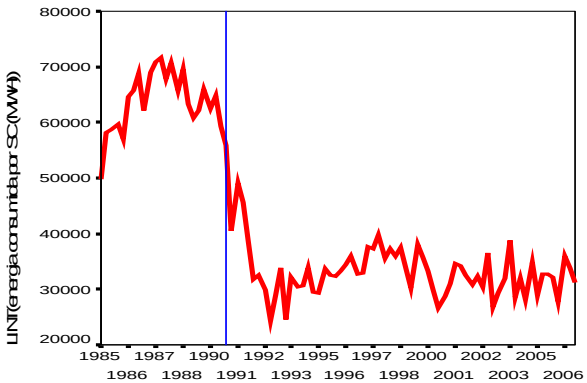
O sea, se indican años/trimestre y el año de comienzo de la misma. Probablemente este gráfico podría ser suficiente para darnos cuenta que - fuera del período especial 1990 a 1994- entre el año 1999 y el 2001 existe un comportamiento irregular, razón que está afectando considerablemente el valor de la varianza, estos valores ocurren en el trimestre Q1-1999 (se consumió 43159.10), en el trimestre Q2-2000 (se consumió 45750.40), en el trimestre Q4-2000 (se consumió 45455) y en el trimestre Q4-2001 (se consumió 41827.90), en los cuales lo más probable es que se deba a errores humanos, (es poco probable que en estos trimestres de estos años el consumo de energía eléctrica en ese municipio se haya elevado tanto). Entonces se tiene para estos casos la alternativa de sustituir los outliers por cierta "interpolación" de valores de sus entornos como se trató en el capítulo 1.

Se ejemplifica el uso del comando especial RMV para tratar valores perdidos. Este comando define una nueva serie (en nuestro caso se nombrará (santa._1)) a partir de una serie existente santa..e el cual presenta cuatro valores eliminados en Q1-1999, Q2-2000, Q4-2000, Q4-2001. Estos valores son sustituidos por algún criterio especificado y en este caso se opta por el criterio LINT (linear interpolation). Este comando se activa con las opciones **Transformar \ remplazar valores perdidos.**

```
RMV
  /santa._1=LINT(santa..e).
```

Result Variable	Missing Values Replaced	First Non-Miss	Last Non-Miss	Valid Cases	Creating Function
SANTA._1	4	1	88	88	LINT(SANTA..E)

Gracias a esta sustitución el gráfico de la serie mejora ostensiblemente



Una vez arreglado hasta cierto punto el problema de la varianza de la serie, se desea ahora eliminar el efecto de la componente estacional (modelo aditivo) en la nueva serie, y para ello hay que desestacionalizar restando a la nueva serie el coeficiente estacional de cada trimestre. Para estimar los coeficientes estacionales, se usa el comando de descomposición estacional del SPSS, el cual se activa con las opciones **Analizar \ Series temporales \ descomposición estacional.**

```
* Seasonal Decomposition.
TSET PRINT=BRIEF NEWVAR=ALL .
SEASON
  /VARIABLES=santa._1
  /MODEL=ADDITIVE
  /MA=EQUAL.
```

El SPSS reporta:

```
Additive Model.  Equal weighted MA method.  Period = 4.
                 Seasonal
      Period        index
         1       -2318.30
         2        790.104
         3        250.108
         4       1278.088
```

Allí se muestran los coeficientes de estacionalidad de cada trimestre y las nuevas variables creadas. Note que la suma de los coeficientes es cero. Las nuevas series creadas son:

SAS (Seasonal Ajusted Series): es la serie desestacionalizada, que se obtiene restando a cada valor original el coeficiente de estacionalidad trimestral (X_t-S_t).

SAF (Seasonal Factors): son los coeficientes de estacionalidad de cada trimestre; o sea la serie estacional pura S_t formada por los coeficientes que se repiten periódicamente, así el primer valor, el quinto, el noveno y así sucesivamente son todos iguales y corresponden al valor -2318.30 mientras que el segundo, el sexto, el décimo etc. corresponden al valor 790.104; en el gráfico se ve claramente esta periodicidad pura. **Ver anexo II-1.**

ERR: corresponde al término de error E_t (diferencia entre la serie SAS_1 y STC_1)**.**

STC (Smothed Trend-Cycle): es la serie de tendencias (tendencias a largo plazo + ciclos**)** re-estimada a partir de la serie ajustada estacionalmente.

Así, en una muestra parcial como la siguiente, de la tabla que reporta el SPSS:

	santa._1	err_1	sas_1	saf_1	stc_1
1	49854.40	-1744.70	52172.70	-2318.30	53917.40
2	58076.20	1305.367	57286.10	790.1036	55980.73
3	58733.50	878.6595	58483.39	250.1081	57604.73
4	59692.00	-709.723	58413.91	1278.088	59123.63
5	57101.40	-1316.18	59419.70	-2318.30	60735.88
6	64568.80	761.5079	63778.70	790.1036	63017.19
7	65745.30	718.1595	65495.19	250.1081	64777.03
8	68853.00	1522.377	67574.91	1278.088	66052.53

se tiene que :

SAS_1=(Santa..1)-(SAF_1) o bien $X_t - S_t = T_t + C_t + E_t$

ERR_1 =(SAS_1) -(STC_1)

STC_1 =(Santa..1) –(SAF_1)-(ERR_1) o bien $X_t - S_t - E_t = T_t + C_t$

Para obtener los índices estacionales (multiplicativos) se procede en el SPSS de forma similar a la de obtener los coeficientes estacionales, con la salvedad de que hay que activar el modelo multiplicativo en vez del modelo aditivo:

```
* Seasonal Decomposition.
TSET PRINT=BRIEF NEWVAR=ALL .
SEASON
  /VARIABLES=santa._1
  /MODEL=MULTIPLICATIVE
  /MA=EQUAL.
```

Se obtiene la siguiente salida:

```
Results of SEASON procedure for variable SANTA._1
Multiplicative Model.  Equal weighted MA method.  Period = 4.
                  Seasonal
                     index
      Period      (* 100)
         1         94.612
         2        101.832
         3        100.669
         4        102.888
```

Allí, se muestran los índices de estacionalidad de cada trimestre y las nuevas variables creadas. Note que en este caso el promedio de los índices es igual a 100. Las variables que se generan son:

SAF (<u>Seasonal Factors</u>): son los índices de estacionalidad de cada trimestre; o sea la serie estacional pura S_t formada por los coeficientes que se repiten periódicamente. El nombre de índice en lugar de coeficiente se debe a su naturaleza multiplicativa ya que muestra el cambio relativo de la serie introducida por la estacionalidad.

SAS (Seasonal Ajusted Series): es la serie desestacionalizada, que se obtiene mediante el cociente $X_t : S_t = T_t \cdot C_t \cdot E_t$.

ERR: corresponde al término de error E_t (cociente entre la serie SAS_1 y STC_1).

STC(Smmothed Trend-Cycle):es la serie de tendencias(tendencias a largo plazo + ciclos**)** re-estimada a partir de la serie ajustada estacionalmente.

Así, en una muestra parcial como la siguiente, de la tabla que reporta el SPSS:

	santa._1	err_2	sas_2	saf_2	stc_2
1	49854.40	.97425	52693.68	.94612	54086.65
2	58076.20	1.01801	57031.59	1.01832	56022.86
3	58733.50	1.01335	58343.31	1.00669	57574.81
4	59692.00	.98170	58016.57	1.02888	59098.01
5	57101.40	.99200	60353.41	.94612	60840.00
6	64568.80	1.00739	63407.40	1.01832	62942.57
7	65745.30	1.00900	65308.53	1.00669	64725.75
8	68853.00	1.01438	66920.44	1.02888	65971.79

SAS_2 = (Santa..1)/ (SAF_2) o sea $X_t : S_t = T_t \cdot C_t \cdot E_t$

ERR_2= (SAS_2)/ (STC_2)

STC_2= (Santa..1)/ (SAF_2**) (**ERR_2) o sea $X_t : (S_t \cdot E_t) = T_t \cdot C_t$

2.6.2 Estimación de la tendencia en la serie Santa Clara.

Ya en las variables SAS_1 y SAS_2, se tienen las series desestacionalizadas. Estas son las series que tomamos para determinar la tendencia, a la cual se puede ajustar el mejor modelo (en realidad, también se pueden tomar las variables STC_1 y STC_2).

Para continuar el análisis, en el SPSS se elige la opción regresión y en particular la estimación curvilínea, que representa la ventaja de considerar varios modelos. Como variable dependiente se introduce a SAS_1 y como variable independiente el tiempo.

```
* Curve Estimation.
PREDICT THRU END.
CURVEFIT /VARIABLES=sas_1
  /CONSTANT
  /MODEL= QUADRATIC
  /PRINT ANOVA
  /PLOT FIT
  /SAVE=PRED RESID .
```

Se elige la opción cuadrática porque fue la que mejor se aproximaba como tendencia.

```
Dependent variable.. SAS_1              Method.. QUADRATI
Listwise Deletion of Missing Data
Multiple R           .85537
R Square             .73165
Adjusted R Square    .72534
Standard Error   7525.96784

            Analysis of Variance:
               DF   Sum of Squares      Mean Square
Regression      2    13126705194.0     6563352597.0
Residuals      85     4814416317.8       56640192.0
F =     115.87801       Signif F =  .0000
-------------------- Variables in the Equation --------------------
Variable               B        SE B       Beta         T        Sig T
Time           -1264.371991  127.710378  -2.249342    -9.900     .0000
Time**2            9.361142    1.390376   1.529690     6.733     .0000
(Constant)     73093.772174 2462.568089               29.682     .0000
```

En este caso ERR_3 representa la serie diferencia entre la serie desestacionalizada SAS_1 y la tendencia cuadrática según el modelo aditivo FIT_3 (Sas_1-Fit_3).

Como se puede apreciar, todos los parámetros son significativos y el modelo

$\mathbf{T_t = 73093.77 - 1264.37\ t + 9.36\ t^2}$ es significativo aunque sólo explica el 73.16% de las variaciones totales; el ploteo de ambos gráficos, de la tendencia y de la serie desestacionalizada SAS_1 se puede ver en el **anexo II-2.**

Para continuar el análisis se realiza nuevamente el ajuste pero esta vez según el modelo multiplicativo y usando la serie desestacionalizada SAS_2. La opción cuadrática vuelve a ser la mejor.

```
* Curve Estimation.
PREDICT THRU END.
CURVEFIT /VARIABLES=sas_2
  /CONSTANT
  /MODEL=QUADRATIC
  /PRINT ANOVA
  /PLOT FIT
  /SAVE=PRED RESID .
```

El SPSS reporta:

```
Dependent variable.. SAS_2              Method.. QUADRATI
Listwise Deletion of Missing Data
Multiple R           .85613
R Square             .73295
Adjusted R Square    .72667
Standard Error   7503.68784
            Analysis of Variance:
               DF   Sum of Squares      Mean Square
Regression      2    13135896543.1     6567948271.5
Residuals      85     4785953151.7       56305331.2
F =     116.64878       Signif F =  .0000
-------------------- Variables in the Equation --------------------
```

Variable	B	SE B	Beta	T	Sig T
Time	-1264.724395	127.332302	-2.251178	-9.932	.0000
Time**2	9.363259	1.386260	1.530858	6.754	.0000
(Constant)	73108.558533	2455.277861		29.776	.0000

Aquí de forma similar, ERR_4 representa la serie diferencia entre la serie desestacionalizada SAS_2 y la tendencia cuadrática según el modelo multiplicativo FIT_4.

Según el reporte $\mathbf{T_t = 73108.55-1264.72\ t + 9.36\ t^2}$ es significativo y explica el 73.29% de las variaciones totales. Además todos los parámetros son significativos.

Al comparar ambas tendencias, se observa que ambas son cuadráticas, gráficamente similares pero con el modelo multiplicativo R^2 es ligeramente mayor (0.73295 contra 0.73165) **ver anexo II-2.**

En el reporte que brinda el SPSS, la serie FIT_3 representa el estimado del ajuste de la serie desestacionalizada según el modelo aditivo, y la serie Fit_4 representa el estimado del ajuste de la serie desestacionalizada según el modelo multiplicativo. En ambos casos se usó como se ha explicado, la regresión con el modelo cuadrático, a través de la opción curvilínea.

2.6.3 Modelos hallados para la serie de Santa Clara.

Después de estimada la tendencia, el paso siguiente es obtener una serie sin tendencias y sin estacionalidad para estudiar las componentes cíclicas y aleatorias; la forma de lograr esto depende del modelo adoptado: si el modelo es aditivo mediante la diferencia y si el modelo es multiplicativo mediante la división, es decir:

$\mathbf{Y_t = T_t+S_t+C_t+E_t}$ **Modelo aditivo**

$\mathbf{(Y_t - Tt-St) = C_t+E_t}$

$\mathbf{Y_t = T_t \,.\, S_t \,.\, C_t \,.\, E_t}$ **Modelo multiplicativo**

$\mathbf{(Y_t/(T_t \,.\, S_t)) = C_t \,.\, E_t}$

En muchos casos como este, no resulta de interés realizar un estudio de la componente cíclica, en particular cuando se desea pronosticar a corto plazo (tal es el objetivo) la variable estudiada o cuando existe alguna razón para suponer que no actúan factores sobre esta variable que provoquen un comportamiento cíclico en ella. En conclusión, los modelos estimados son los siguientes:

$$\mathbf{Y_t=T_t+S_t+E_t=}\begin{cases} 73093.77 - 1264.37\,t + 9.36\,t^2 - 2318.30 + E_t & si\ t\varepsilon\ I\ trimestre \\ 73093.77 - 1264.37\,t + 9.36\,t^2 + 790.104 + E_t & si\ t\varepsilon\ II\ trimestre \\ 73093.77 - 1264.37\,t + 9.36\,t^2 + 250.108 + E_t & si\ t\varepsilon\ III\ trimestre \\ 73093.77 - 1264.37\,t + 9.36\,t^2 + 1278.088 + E_t & si\ t\varepsilon\ IV\ trimestre \end{cases}$$

que se obtiene mediante la suma de la tendencia estimada según el modelo aditivo, el coeficiente estacional y la serie de perturbaciones aleatorias o.

$$\mathbf{Y_t=T_t*S_t*E_t=}\begin{cases}(73108.55-1264.72t+9.36t^2)(0.946)E_t & si\ t\varepsilon\ I\ trimestre\\ (73108.55-1264.72t+9.36t^2)(1.018)E_t & si\ t\varepsilon\ II\ trimestre\\ (73108.55-1264.72t+9.36t^2)(1.006)E_t & si\ t\varepsilon\ III\ trimestre\\ (73108.55-1264.72t+9.36t^2)(1.028)E_t & si\ t\varepsilon\ IV\ trimestre\end{cases}$$

que se obtiene mediante el producto de la tendencia estimada según el modelo multiplicativo, el índice estacional y la serie de perturbaciones aleatorias.

Así, si se quisiera hacer un pronóstico para el año 2006 primer trimestre, tan sólo habría que efectuar un estimado con FIT_3 (SAS_1) hasta esa fecha (con ayuda del SPSS) y luego sumarle el coeficiente estacional SAF_1**,** para el modelo aditivo**,** y para el modelo multiplicativo, hacer un estimado con FIT_4 (SAS_2) hasta esa fecha y luego efectuar el producto con SAF_2. El estimado se logra con:

```
* Curve Estimation.
PREDICT THRU year 2006 quarter 1.
CURVEFIT /VARIABLES=sas_1
  /CONSTANT
  /MODEL= QUADRATIC
    /PLOT FIT
  /SAVE=PRED.
```

$Y_t=T_t+S_t+E_t$ = 33256.401-2318.30+ E_t=30938.1+ E_t

El cual comparado con el valor real para Santa Clara en este trimestre que fue de **<u>27791</u> MWH,** no constituye un excelente pronóstico.

De forma similar se ejecuta el comando:

```
* Curve Estimation.
PREDICT THRU year 2006 quarter 1.
CURVEFIT /VARIABLES=sas_2
  /CONSTANT
  /MODEL= QUADRATIC
    /PLOT FIT
  /SAVE=PRED.
```

$Y_t=T_t \cdot S_t \cdot E_t$=(33256.53)(0.946) E_t=(31460.68) E_t el cual es un estimado más alejado que el anterior.

Modelo	Año	Trimestre	Pronóstico	Valor real
Aditivo	**2006**	**1**	**30938.1 Mwh**	**27791Mwh**
Multiplicativo	**2006**	**1**	**31460.68 Mwh**	**27791Mwh**

Pero aquí existe un problema: los modelos anteriores no son válidos salvo que se pruebe que los $\mathbf{E_t}$ constituyen al menos una serie ARMA (lo ideal fuese que constituyeran un ruido blanco). Por tanto, esa necesario calcular las series de errores $\mathbf{E_t}$ según los modelos hallados, lo cual se logra con las opciones: **Transformar/calcular** o procediendo a nivel de sintaxis de la siguiente forma:

```
COMPUTE pronos.a = saf_1+fit_3 .
EXECUTE.
COMPUTE error.a =santa..1-pronos.a .
EXECUTE.
COMPUTE pronos.c = saf_2 * fit_4.
EXECUTE.
COMPUTE error.c =santa..1-pronos.c .
EXECUTE.
```

Para probar que la serie de residuales se comporta como una serie ARMA, se requiere lo expuesto en el capítulo 1 y en cierto sentido los ejemplos del capítulo 3 (esto supone un salto en la exposición del presente trabajo pero es imprescindible), **ver anexoII-3.**

Como la ACF declina sinusoidalmente a 0 mientras que la PACF muestra 1 espiga significativa, se intuye por tanto que se trata de un modelo autorregresivo de orden 1, esto es un modelo ARMA(1 0).Esta opción la podemos encontrar a través de **Analizar \ series temporales \ Arima.**

```
ARIMA error.a
  /MODEL=( 1 0 0 )( 0 0 0 ) NOCONSTANT
  /MXITER 10
  /PAREPS .001
  /SSQPCT .001
  /FORECAST EXACT.
```

El reporte del SPSS es:

```
Split group number: 1  Series length: 88
FINAL PARAMETERS:
Number of residuals  88
Standard error       3848.9105
Log likelihood       -851.60184
AIC                  1705.2037
SBC                  1707.681
```

```
 Analysis of Variance:
               DF  Adj.        Sum of Squares    Residual Variance
Residuals      87              1310967597.3           14814111.8
          Variables in the Model:
                B         SEB        T-RATIO          APPROX. PROB.
AR1          .88125696   .04872020   18.088122         .0000000
```

Se observa que el parámetro AR1 es significativo y los residuales de este modelo se comportaron adecuadamente (los resultados se muestran en forma parcial):

```
ACF
 VARIABLES= err_7
 /NOLOG
 /MXAUTO 16
 /SERROR=IND
 /PACF.
```

```
Autocorrelations:   ERR_7   Error for ERROR.A from ARIMA, MOD_10 NOC
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                   ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  1  -.185   .105                ****⇔    .                       3.104   .078
  2  -.018   .104                .   *   .                       3.135   .209
  3   .085   .104                .   ⇔** .                        3.809   .283
  4   .093   .103                .   ⇔** .                        4.623   .328
  5  -.047   .102                .  *⇔    .                       4.834   .436
  6   .077   .102                .   ⇔** .                        5.401   .493
  7  -.004   .101                .   *   .                       5.403   .611
  8   .033   .101                .   ⇔*   .                       5.508   .702
  9  -.169   .100                .***⇔    .                       8.382   .496
 10   .096   .099                .   ⇔** .                        9.324   .502
 11  -.027   .099                .  *⇔    .                       9.401   .585
 12  -.117   .098                . **⇔    .                      10.821   .544
 13  -.054   .097                .  *⇔    .                      11.125   .600
 14  -.007   .097                .   *   .                      11.130   .676
 15  -.050   .096                .  *⇔    .                      11.401   .724
 16  -.007   .095                .   *   .                      11.406   .784
 Partial Autocorrelations:   ERR_7   Error for ERROR.A from ARIMA, MOD_10 NOC
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                   ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  1  -.185   .107                ****⇔    .
  2  -.054   .107                .  *⇔    .
  3   .074   .107                .   ⇔*   .
  4   .127   .107                .   ⇔***.
  5   .000   .107                .   *   .
  6   .068   .107                .   ⇔*   .
  7   .003   .107                .   *   .
  8   .030   .107                .   ⇔*   .
  9  -.176   .107                ****⇔    .
 10   .018   .107                .   *   .
```

Aquí, todas las autocorrelaciones quedan dentro de la banda de confianza alrededor de cero, y por tanto no resultan significativas. El test de Box-Ljung en cada valor de h informa que no existen razones para sospechar que dichas autocorrelaciones difieran de las correspondientes a un ruido blanco. Por tanto con estos resultados se garantiza que la serie de residuales $\mathbf{E_t}$ corresponde a una serie ARMA y por tanto el modelo aditivo hallado es

válido. El ploteo tanto de la serie tratada para Santa Clara como la serie pronóstico con el modelo aditivo se muestra en el **anexo II-4.**

De forma similar se procede para el modelo multiplicativo, en el cual también se puede mostrar que los residuales cumplen el requisito de tener estructura ARMA.

2.6.4. Comparación de series pronósticos para la serie Santa Clara.

Se pudiera cuestionar el hecho de no haberse trabajado con las series stc_1 y stc-2, pero este trabajo se realizó, y de hecho algunos resultados parciales se muestran a continuación.

1ro) Se calcularon las cuatro tendencias para las cuatro series (sas_1, stc_1, sas_2 y stc_2) dos para el modelo aditivo y dos para el modelo multiplicativo y como se muestra en la siguiente tabla, el ajuste cuadrático para la serie stc_2 fue el mejor:

Sas_1	Stc_1	Sas_2	Stc_2
Multiple R .85537	Multiple R .87103	Multiple R .85613	Multiple R .87130
R Square .73165	R Square .75870	R Square .73295	R Square **.75916**
Adjusted R Sq. .72534	Adjusted R Sq .75302	Adjusted R Sq. .72667	Adjusted R Sq .75349
Std. Error 7525.96784	Std Error 991.21877	Std Error 7503.68784	Std Error 984.03036

2do) Se calcularon las cuatro series pronósticos y los residuales: **Ver anexo II-5**

3ro) Se realizó una comparación con el comando FIT, esto sólo se puede hacer a nivel de sintaxis de comandos en el SPSS.

```
FIT
ERRORS= error.a error.b error.c error.d /
OBS= santa..1 santa..1 santa..1 santa..1/
DFE= 84 84 84 84.
```

FIT Error Statistics

Error Variable		ERROR.A	ERROR.B	ERROR.C	ERROR.D
Observed Variable		SANTA..1	SANTA..1	SANTA..1	SANTA..1
N of Cases	Use	88	88	88	88
Deg Freedom	Use	84	84	84	84
Mean Error	Use	.0000	-.8832	8.2142	5.2503
Mean Abs Error	Use	5769.9551	5773.2567	5742.3706	5746.2551
Mean Pct Error	Use	-2.9397	-2.9601	-2.9328	-2.9595
Mean Abs Pct Err	Use	14.2097	14.2159	14.1435	14.1514
SSE	Use	4814416318	4814514645	4762139087	4762316036
MSE	Use	57314480.0	57315650.5	56692132.0	56694238.5
RMS	Use	7570.6327	7570.7100	7529.4178	7529.5577
Durbin-Watson	Use	.2777	.2777	.2706	.2705

El comando Fit permite obtener algunos estadísticos interesantes que son especialmente útiles para comparar varios modelos posibles.

Los resultados de este comando son en orden de aparición:

- **Error Variable: variable que se analiza.**

- **Observed variable: Variable que se utiliza como base en la comparación.**
- **N of Cases: Número de casos en los períodos de uso y validación (se muestra en el capítulo 3).**
- **Deg Freedom: Grados de libertad.**
- **Mean Error: Error medio.**
- **Mean Abs Error: Error medio absoluto, da el valor medio del error en valor absoluto.**
- **Mean Pct error: Errores en porcientos, se calculan utilizando como denominador los valores observados de la serie y luego se promedian incluyendo signos.**
- **Mean Abs Pct Error: Similar al anterior, sólo que los valores se promedian en valor absoluto.**
- **SSE: Suma de cuadrado de los errores, esto es, la suma de los cuadrados de las diferencias entre los valores observados de la serie y los predichos por el modelo.**
- **MSE: Es la media de la SSE, esto es la SSE dividida por los grados de libertad del error.**

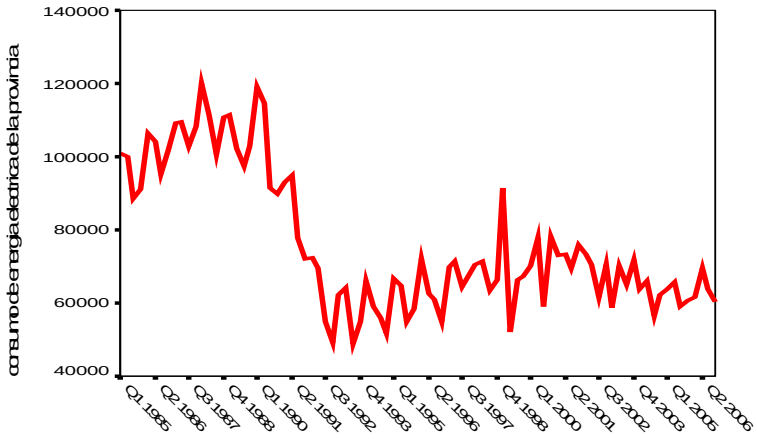
Si se utiliza FIT para comparar dos modelos, el criterio más fuerte de comparación se formula sobre la base de minimizar este último estadístico.

- **RMS: Es la raíz cuadrada de la MSE, permite tener un estadístico en el mismo sistema de unidades que la serie observada y la serie de errores.**
- **El test de Durbin- Watson verifica la hipótesis nula de que los residuales de la regresión son independientes, contra la hipótesis alternativa de que siguen un proceso autorregresivo de primer orden; su valor se encuentra entre 0 y 4. Un valor cercano a 2 indica poca autocorrelación y es lo deseado.**

Así, en el estudio presente, con la serie de Santa Clara, se puede observar que estadísticos como el MAPE y el RMS son mejores en el modelo multiplicativo, no obstante presentar peores valores en el indicador de Durbin-Watson. Por tanto, hay aspectos a favor y en contra entre los modelos aditivo y multiplicativo, y se mantiene la estimación realizada con los modelos a partir de las series Sas_1 y Sas_2.

2.7. Serie consumo eléctrico en la provincia Villa Clara.

Se elige esta serie por dos motivos: primero por tratarse de una serie provincial y segundo por tratarse de una serie "más complicada", con tendencia decreciente, entre otros aspectos, como se puede apreciar en el siguiente gráfico.



En esta serie se pretende también lograr modelos según el enfoque clásico, pero todo de una manera breve teniendo en cuenta que un estudio en detalle se realizó con la serie anterior.

Definidas las variables año y mes con el comando DATE, se plotea la serie y luego se utiliza el comando SEASON para separar la componente estacional con la opción modelo aditivo, todo lo cual se logra con:

```
* Seasonal Decomposition.
TSET PRINT=BRIEF NEWVAR=ALL .
SEASON
  /VARIABLES=electric
  /MODEL=ADDITIVE
  /MA=EQUAL.
```

La salida de este comando es:

```
Results of SEASON procedure for variable ELECTRIC
Additive Model.  Equal weighted MA method.  Period = 4.
                  Seasonal
      Period         index
         1        4639.629
         2        2602.506
         3        -5433.40
         4        -1808.74
```

Una vez separada la tendencia estacional, se utiliza la serie desestacionalizada SAS_1 para investigar por regresión si existe efectivamente una tendencia significativa. Ello se logra tal como se hizo en el epígrafe anterior mediante el comando:

```
* Curve Estimation.
PREDICT THRU END.
```

```
CURVEFIT /VARIABLES=sas_1
  /CONSTANT
  /MODEL=QUADRATIC
  /PRINT ANOVA
  /PLOT FIT
  /SAVE=PRED .
```

```
Dependent variable.. SAS_1              Method.. QUADRATI
Listwise Deletion of Missing Data

Multiple R           .79405
R Square             .63051
Adjusted R Square    .62182
Standard Error  11444.03392
            Analysis of Variance:
               DF   Sum of Squares      Mean Square
Regression      2    18996303281.7     9498151640.8
Residuals      85    11132102548.6      130965912.3
F =      72.52385       Signif F =  .0000
-------------------- Variables in the Equation --------------------
Variable                   B        SE B       Beta          T  Sig T
Time           -1577.865385  194.197202  -2.166143    -8.125  .0000
Time**2           11.996413    2.114214   1.512734     5.674  .0000
(Constant)    115155.361532 3744.596486                30.752  .0000
```

Ahora, la tendencia $\mathbf{T_t=115155.36-1577.865\ t\ +11.996\ t^2}$ es significativa y explica el 63.051% de las variaciones totales. Además todos los parámetros son significativos.

Luego de estimar la tendencia, sólo resta comprobar si los residuales $\mathbf{E_t}$ conforman una serie ARMA para poder hablar de la existencia de un modelo aditivo para esta serie de la forma

$\mathbf{Y_t= T_t+S_t+E_t.}$Los residuales se calculan mediante:

```
COMPUTE pronos.a=fit_2+saf_1.
EXECUTE.
COMPUTE error.a = electric-pronos.a.
EXECUTE.
```

El próximo paso es reconocer, mediante los correlogramas de **error.a** , cuál es la estructura de esta serie, **ver anexo II-6.**

Como se puede observar la ACF declina sinusoidalmente a 0 mientras que la PACF muestra 2 espigas significativas. Se intuye por tanto que se trata de un modelo autorregresivo de orden 2, esto es un modelo ARMA(2,0).

```
ARIMA error.a
  /MODEL=( 2 0 0 )( 0 0 0 ) NOCONSTANT
  /MXITER 10
  /PAREPS .001
  /SSQPCT .001
  /FORECAST EXACT.
```

```
Split group number: 1  Series length: 88
FINAL PARAMETERS:
Number of residuals  88
Standard error       6958.9058
Log likelihood       -903.05673
AIC                  1810.1135
SBC                  1815.0681

              Analysis of Variance:

              DF  Adj. Sum of Squares    Residual Variance
Residuals     86         4221396720.3           48426369.4
           Variables in the Model:
                B          SEB      T-RATIO   APPROX. PROB.
AR1    .43968145   .09731962   4.5179115       .00001977
AR2    .41192480   .09783017   4.2106109       .00006243
```

Se observa que ambos parámetros AR1 y AR2 son significativos, pero además, algo muy importante, los residuales se comportan como un ruido blanco, lo cual se verifica mediante el reporte:

```
ACF
 VARIABLES= err_3
 /NOLOG
 /MXAUTO 16
 /SERROR=IND
 /PACF.
```

```
Autocorrelations:   ERR_3   Error for ERROR.A from ARIMA, MOD_7 NOCO
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                   ¤⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩¤
  1  -.059   .105                .  *⇔   .                          .312   .577
  2  -.067   .104                .  *⇔   .                          .726   .696
  3   .102   .104                .   ⇔** .                         1.702   .636
  4   .117   .103                .   ⇔** .                         2.989   .560
  5   .155   .102                .   ⇔***.                         5.292   .381
  6  -.114   .102                . **⇔   .                         6.557   .364
  7   .082   .101                .   ⇔** .                         7.218   .407
  8   .155   .101                .   ⇔***.                         9.597   .294
  9  -.196   .100                ****⇔   .                        13.463   .143
 10  -.035   .099                .  *⇔   .                        13.587   .193
 11  -.019   .099                .   *   .                       13.623   .255
 12   .066   .098                .   ⇔*  .                        14.083   .295
 13  -.142   .097                .***⇔   .                        16.223   .237
 14  -.107   .097                . **⇔   .                        17.455   .233
 15   .126   .096                .   ⇔***.                        19.190   .205
 16  -.143   .095                .***⇔   .                        21.447   .162
Partial Autocorrelations:   ERR_3   Error for ERROR.A from ARIMA, MOD_7 NOCO
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                   ¤⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩¤
  1  -.059   .107                .  *⇔   .
  2  -.071   .107                .  *⇔   .
  3   .095   .107                .   ⇔** .
  4   .126   .107                .   ⇔***.
  5   .190   .107                .   ⇔****
  6  -.088   .107                . **⇔   .
  7   .068   .107                .   ⇔*  .
  8   .108   .107                .   ⇔** .
  9  -.205   .107                ****⇔   .
 10  -.073   .107                .  *⇔   .
 11  -.067   .107                .  *⇔   .
 12   .031   .107                .   ⇔*  .
```

```
 13  -.130   .107                .***⇔   .
 14  -.017   .107                .   *   .
 15   .087   .107                .   ⇔** .
 16  -.123   .107                . **⇔   .
```

Por tanto, se puede formular el modelo aditivo de la forma:

$$\mathbf{Y_t= T_t+S_t+E_t=}\begin{cases}115155.36-1577.86t+11.996t^2+4639.629 & si\ t\varepsilon\ I\ trimestre\\ 115155.36-1577.86t+11.996t^2+2602.506 & si\ t\varepsilon\ II\ trimestre\\ 115155.36-1577.86t+11.996t^2-5433.40 & si\ t\varepsilon\ III\ trimestre\\ 115155.36-1577.86t+11.996t^2-1808.74 & si\ t\varepsilon\ IV\ trimestre\end{cases}$$

Con este resultado, se termina este capítulo dedicado al trabajo con los modelos clásicos de series de tiempo.

Conclusiones parciales

En este capítulo se muestra, de manera resumida, los conceptos fundamentales de la modelación de series temporales según el enfoque clásico. Se hallan los modelos matemáticos de las series de consumo eléctrico de Santa Clara y de la provincia. Se realizan pronósticos con esta concepción. Se muestra la bondad de los cálculos desde el punto de vista computacional pero también la posible insuficiencia en el pronóstico.

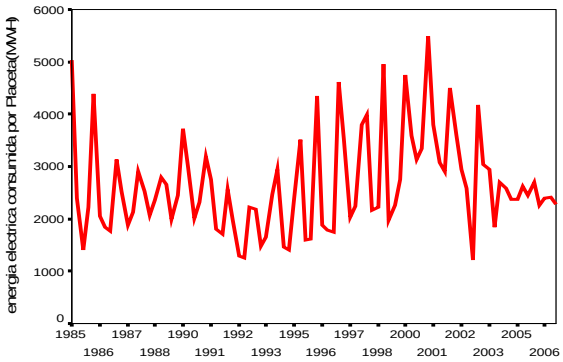
Capítulo 3: Análisis de series por modelación Arima

En este capítulo se hallan modelos matemáticos del tipo ARIMA para algunas series de la provincia Villa Clara y de los diferentes municipios en los rubros consumo de energía eléctrica y petróleo, específicamente la serie de consumo energía eléctrica de Placetas, y además se tratará la serie de consumo energía eléctrica de Santa Clara, con el propósito de establecer una comparación con el modelo hallado según el enfoque clásico, así como se muestran pronósticos en base a los mismos. Se sigue la metodología de Box-Jenkins para el análisis de series de tiempo y se trabaja también con el software estadístico SPSS.

La tabla original de datos suministrados por la Oficina Nacional de Estadística de Villa Clara se muestra parcialmente en el **anexo III-1** y de ahí, previa organización de los datos en columnas, se formaron las diferentes series provinciales y municipales en los rubros de consumo de energía eléctrica y de petróleo. A continuación se muestran ejemplos de cómo se obtuvieron los modelos matemáticos.

3.1. Serie consumo de energía eléctrica del municipio de Placetas.

Esta serie tiene longitud de 88 observaciones del tipo trimestral, cuatro por año, y su gráfico es:



En el mismo se percibe al menos que no hay una evidente tendencia lineal, ni hay síntomas claros de heteroscedasticidad, aunque sí una aparente tendencia a la periodicidad. La serie original se nombrará **place..e.**

Corresponde, según la metodología, hacer ahora el cálculo y ploteo de las funciones de autocorrelación y autocorrelación parcial. Esto se logra para los 88 casos de la muestra con el comando:

```
ACF
  VARIABLES= place..e
  /NOLOG
  /MXAUTO 16
  /SERROR=IND
  /PACF.
```

Cuya salida se muestra a continuación:

```
Autocorrelations:   PLACE..E   energia electrica consumida por Placeta(
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1  Box-Ljung  Prob.
                   ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  1   .180   .105                .   ⇔****                       2.958   .085
  2  -.146   .104                .***⇔   .                       4.927   .085
  3   .155   .104                .   ⇔***.                       7.174   .067
  4   .657   .103                .   ⇔***.*********             47.896   .000
  5   .208   .102                .   ⇔****                      52.035   .000
  6  -.172   .102                .***⇔   .                      54.900   .000
  7   .099   .101                .   ⇔** .                      55.860   .000
  8   .432   .101                .   ⇔***.*****                 74.333   .000
  9   .103   .100                .   ⇔** .                      75.405   .000
 10  -.210   .099                ****⇔   .                      79.883   .000
 11   .029   .099                .   ⇔*  .                      79.969   .000
 12   .363   .098                .   ⇔***.***                   93.716   .000
 13   .022   .097                .   *   .                     93.767   .000
 14  -.242   .097               *.***⇔   .                     100.030   .000
 15   .039   .096                .   ⇔*  .                     100.193   .000
 16   .212   .095                .   ⇔****                     105.123   .000
Total cases:  88     Computable first lags:  87
Partial Autocorrelations:   PLACE..E   energia electrica consumida por Placeta(
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                   ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  1   .180   .107                .   ⇔****
  2  -.185   .107                ****⇔   .
  3   .235   .107                .   ⇔***.*
  4   .612   .107                .   ⇔***.********
  5   .089   .107                .   ⇔** .
  6  -.112   .107                . **⇔   .
  7  -.006   .107                .   *   .
  8  -.036   .107                .  *⇔   .
  9  -.137   .107                .***⇔   .
 10  -.101   .107                . **⇔   .
 11  -.019   .107                .   *   .
 12   .150   .107                .   ⇔***.
 13  -.039   .107                .  *⇔   .
 14  -.031   .107                .  *⇔   .
 15   .095   .107                .   ⇔** .
 16  -.153   .107                .***⇔   .
```

Para ambas funciones de autocorrelación se muestra el valor de la autocorrelación para cada retardo “h” desde 1 hasta 16 (esto es modificable), sus errores estándar, y con una línea de puntos, el intervalo de confianza fuera del cual se puede considerar que la autocorrelación es significativamente diferente de cero. En el caso de la ACF, se muestra en cada retardo “h” el valor del estadístico de Box-Ljung y su significación. Este estadístico sirve para verificar la hipótesis nula que un conjunto de autocorrelaciones muestrales esta asociada con una serie aleatoria; más precisamente, que las autocorrelaciones en cada retardo se corresponde con la que podría tener un ruido blanco para ese retardo.

Los correlogramas ratifican que la serie no es estacionaria, y no sólo eso, sino que resulta evidente la tendencia periódica anual de la serie y por tanto la estacionalidad S=4; ello se muestra en la ACF con la cantidad de espigas que regularmente y alternativamente salen fuera del intervalo de confianza, además del estadístico Box-Ljung que indica que existen

razones para sospechar que las autocorrelaciones sí difieren a las correspondientes a un ruido blanco.

Como quiera que para estimar el modelo se utilizará sólo una parte de los datos (en este caso 80), es deseable tener alguna medida de si estos 80 primeros casos reflejan suficientemente bien las autocorrelaciones de la serie original. Ello puede lograrse repitiendo los correlogramas con los 80 primeros casos y con los comandos, los cuales lo podemos encontrar en (**Datos \ Seleccionar casos \ Basándose en los rangos y Gráficos \ serie temporal \ Autocorrelaciones), ver anexo III-2**.

Como se puede ver, el resultado ratifica el comportamiento similar de los correlogramas, para proseguir el análisis y la estimación del modelo, se trazan nuevamente los correlogramas pero esta vez los correlogramas estacionales para identificar la estructura estacional (P, D, Q) del modelo, los cuales, aplicando el comando correspondiente puede ser activado siguiendo el siguiente orden de pasos (**Gráficos \ serie temporal \ Autocorrelaciones \ opciones \ mostrar autocorrelaciones**) resultan:

```
ACF
 VARIABLES= place..e
 /NOLOG
 /MXAUTO 40
 /SERROR=IND
 /SEASONAL
 /PACF.
```

```
Autocorrelations:   PLACE..E   energia electrica consumida por Placeta(
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                  ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  4   .657   .108                .   ⇔***.*********             43.766   .000
  8   .425   .105                .   ⇔***.*****                 67.749   .000
 12   .374   .102                .   ⇔***.***                   86.736   .000
 16   .239   .099                .   ⇔***.*                     98.258   .000
 20   .239   .096                .   ⇔***.*                    112.577   .000
 24   .140   .092                .   ⇔***.                     124.101   .000
 28   .039   .089                .   ⇔*  .                     134.216   .000
 32   .053   .086                 .  ⇔* .                      150.607   .000
 36   .073   .082                 .  ⇔* .                      169.315   .000
 40   .133   .078                 .  ⇔***                      184.572   .000
Total cases:  80     Computable first lags:  79
Partial Autocorrelations:   PLACE..E   energia electrica consumida por Placeta(
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                  ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  4   .612   .112                .   ⇔***.********
  8  -.050   .112                .  *⇔    .
 12   .176   .112                .   ⇔****
 16  -.136   .112                .***⇔    .
 20   .161   .112                .   ⇔***.
 24  -.163   .112                .***⇔    .
 28  -.019   .112                .   *   .
 32  -.020   .112                .   *   .
 36   .113   .112                .   ⇔** .
 40   .083   .112                .   ⇔** .
```

Note que la SACF(h) muestra una declinación aunque no tan rápida a cero y que la SPACF(h) presenta una espiga. Se puede tener la duda de si es necesaria una diferenciación

estacional de orden 1. Si se sigue un principio de parsimonia tratando de evitar la pérdida de los 4 primeros datos, se identifica en principio el modelo como ARIMA (0 0 0)(1 0 0)4.

A continuación se emprende la estimación para estudiar los residuales.

```
ARIMA place..e
 /MODEL=( 0 0 0 )( 1 0 0 ) CONSTANT
 /MXITER 10
 /PAREPS .001
 /SSQPCT .001
 /FORECAST EXACT .
```

```
Split group number: 1  Series length: 80

Initial values:
SAR1       .91899
CONSTANT 2615.566
Marquardt constant = .001
Adjusted sum of squares = 45696653.8
              Iteration History:
  Iteration    Adj. Sum of Squares    Marquardt Constant
          1            40863436.9             .00100000
          2            40802204.1             .00010000
Conclusion of estimation phase.
Estimation terminated at iteration number 3 because:
   Sum of squares decreased by less than .001 percent.
FINAL PARAMETERS:
Number of residuals  80
Standard error       711.33568
Log likelihood       -639.20502
AIC                  1282.41
SBC                  1287.1741
            Analysis of Variance:
              DF  Adj. Sum of Squares    Residual Variance
Residuals     78            40802077.7            505998.45
                    Variables in the Model:
                    B          SEB     T-RATIO   APPROX. PROB.
SAR1             .69691      .07716    9.031925        .0000000
CONSTANT    2652.27245   236.60128   11.209882        .0000000
```

El hecho de que el coeficiente autorregresivo estacional (**0.69691**) tenga un valor cercano a 1 pero inferior a este, es también una expresión de la duda sobre si valdría la pena haber diferenciado estacionalmente la serie.

En esta salida, se puede observar como los dos parámetros; la constante y el coeficiente autorregresivo estacional SAR1 son significativos.

Se emprende ahora el chequeo-diagnóstico del ajuste estacional mediante el ploteo de la serie err_1 y el estudio de los correlogramas estacionales:

```
TSPLOT VARIABLES= err_1
 /NOLOG
 /FORMAT NOFILL NOREFERENCE.
```

La serie de residuales (**anexo III-3**) muestra, **excepto al principio**, un comportamiento normal; o sea no muestra tendencias aparentes. A continuación se precisa con el comando ACF hasta que punto no están correlacionados estacionalmente.

```
ACF
 VARIABLES= err_1
 /NOLOG
 /MXAUTO 40
 /SERROR=IND
```

/SEASONAL
/PACF.

```
Autocorrelations:   ERR_1   Error for PLACE..E from ARIMA, MOD_8 CON
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                   ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇩▫
  4  -.032   .108                 .  *⬄   .                          .537   .970
  8  -.116   .105                 . **⬄   .                         5.125   .744
 12   .167   .102                 .   ⬄***.                         8.567   .739
 16  -.144   .099                 .***⬄   .                        13.122   .664
 20   .151   .096                 .   ⬄***.                        17.835   .598
 24   .001   .092                 .   *   .                        18.938   .755
 28  -.132   .089                 .***⬄   .                        22.240   .770
 32  -.010   .086                  .  *  .                         25.702   .777
 36  -.063   .082                  . *⬄  .                         29.953   .751
 40   .040   .078                  .  ⬄* .                         30.607   .858
Total cases:  80     Computable first lags:  79
Partial Autocorrelations:   ERR_1   Error for PLACE..E from ARIMA, MOD_8 CON
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                   ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇩▫
  4  -.036   .112                 .  *⬄   .
  8  -.153   .112                 .***⬄   .
 12   .155   .112                 .   ⬄***.
 16  -.158   .112                 .***⬄   .
 20   .157   .112                 .   ⬄***.
 24  -.047   .112                 .  *⬄   .
 28  -.044   .112                 .  *⬄   .
 32  -.130   .112                 .***⬄   .
 36  -.041   .112                 .  *⬄   .
 40   .048   .112                 .   ⬄*  .
Plot Symbols:      Autocorrelations *     Two Standard Error Limits .
```

Como se puede observar, prácticamente todas las autocorrelaciones quedan dentro de la banda de confianza alrededor de cero, y por tanto no resultan significativas. El test de Box-Ljung en cada valor de h informa que no existen razones para sospechar que dichas autocorrelaciones difieren de las correspondientes a un ruido blanco con probabilidad de error de un 5%. Tampoco aparecen autocorrelaciones parciales significativas en los residuales.

Una vez considerado que los errores no están correlacionados estacionalmente se ha “aislado” la componente estacional, y se debe pasar a analizar si se requieren componentes regulares en el modelo, esto es, determinar “p” y “q”.

Para ello, se regresa al trazado de los correlogramas regulares (h=1,2,3,….) hasta un máximo que podemos reducir de nuevo a 16.

```
ACF
 VARIABLES= err_1
 /NOLOG
 /MXAUTO 16
 /SERROR=IND
 /PACF.
```

```
 Autocorrelations:   ERR_1   Error for PLACE..E from ARIMA, MOD_8 CON

     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                   ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇩▫
  1   .029   .110                 .   ⬄*  .                          .068   .794
  2   .009   .109                 .   *   .                         .075   .963
  3   .066   .108                 .   ⬄*  .                          .450   .930
  4  -.032   .108                 .  *⬄   .                          .537   .970
```

```
  5   .194   .107                .   ⇔****                    3.825   .575
  6  -.010   .106                .   *   .                   3.834   .699
  7   .025   .105                .   *   .                   3.888   .793
  8  -.116   .105                . **⇔    .                   5.125   .744
  9  -.012   .104                .   *   .                   5.139   .822
 10  -.059   .103                .  *⇔    .                   5.468   .858
 11  -.066   .103                .  *⇔    .                   5.878   .881
 12   .167   .102                .   ⇔***.                    8.567   .739
 13  -.082   .101                . **⇔    .                   9.220   .756
 14  -.018   .100                .   *   .                   9.251   .815
 15   .131   .100                .   ⇔***.                   10.994   .753
 16  -.144   .099                .***⇔    .                  13.122   .664
Total cases:  80     Computable first lags:  79
Partial Autocorrelations:   ERR_1   Error for PLACE..E from ARIMA, MOD_8 CON
 Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                  ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  1   .029   .112                .   ⇔*  .
  2   .009   .112                .   *   .
  3   .066   .112                .   ⇔*  .
  4  -.036   .112                .  *⇔    .
  5   .196   .112                .   ⇔****
  6  -.028   .112                .  *⇔    .
  7   .032   .112                .   ⇔*  .
  8  -.153   .112                .***⇔    .
  9   .019   .112                .   *   .
 10  -.115   .112                . **⇔    .
 11  -.026   .112                .  *⇔    .
 12   .155   .112                .   ⇔***.
 13  -.032   .112                .  *⇔    .
 14  -.012   .112                .   *   .
 15   .157   .112                .   ⇔***.
 16  -.158   .112                .***⇔    .
```

Algo curioso se observa en este caso, los residuales se comportan como no correlacionados según nos informa el estadístico de Box-Ljung en cada valor de h; pues la mayoría de las autocorrelaciones quedan dentro o en la frontera de la banda de confianza alrededor de cero y tampoco aparecen muchas autocorrelaciones parciales significativas en los residuales. Podría pensarse como conseguir mejorar este modelo pero como primera versión se toma el actual tal como está, sin componentes regulares “p”y ”q”.

Un paso complementario en el diagnóstico, es comprobar hasta que punto la serie estimada es capaz de pronosticar bien los valores reservados para la validación (años 2005 y 2006). Para lograr que el SPSS haga esto y al mismo tiempo no tenga que repetir la estimación de los parámetros se utiliza el comando de la forma siguiente:

```
ARIMA place..e
 /MODEL=( 0 0 0 )( 1 0 0 ) CONSTANT
 /MXITER 10
 /PAREPS .001
 /SSQPCT .001
 /FORECAST EXACT .
Applying INITIAL estimates from MOD_12
Split group number: 1  Series length: 88
```

```
FINAL PARAMETERS:

Number of residuals  88
Standard error       681.23523
Log likelihood       -699.28884
AIC                  1402.5777
```

```
SBC                  1407.5324
            Analysis of Variance:
              DF  Adj. Sum of Squares    Residual Variance
Residuals     86           41134675.8            464081.43
                     Variables in the Model:
                     B          SEB     T-RATIO   APPROX. PROB.
SAR1             .69671       .07357    9.470433         .0000000
CONSTANT    2633.10774   217.78098   12.090623         .0000000
```

El chequeo diagnóstico del modelo integral, se inicia ploteando los residuales y sus correlogramas, **ver anexo III-4.**

Observe como se mantiene la no correlación de los residuales según nos informa el test de Box-Ljung que no existen razones para sospechar que dichas autocorrelaciones difieren de las correspondientes a un ruido blanco y tampoco aparecen autocorrelaciones significativas en las parciales. No obstante prosigue el estudio hasta el final antes de mejorar el modelo.

El próximo paso es utilizar el comando FIT, el cual permite, como se vio en el capítulo 2, obtener algunos estadísticos interesantes de diagnóstico y que son especialmente útiles para comparar varios modelos posibles o dentro de un mismo modelo para comparar errores en el período de uso y el de validación. Ello se logra mediante la sintaxis:

```
USE YEAR 1985 QUARTER 1 THRU 2004 QUARTER 4 .
PREDICT YEAR 2005 QUARTER 1 THRU YEAR 2006 QUARTER 4 .
FIT
ERRORS=ERR_2/
OBS= PLACE..E/
DFE= 78.
```

```
                          FIT Error Statistics
 Error Variable                        ERR_2
 Observed Variable                     PLACE..E
 N of Cases          Use                80
                     Predict             8
 Deg Freedom         Use                78
                     Predict             8
 Mean Error          Use                 2.7102
                     Predict           -71.5321
 Mean Abs Error      Use               492.3472
                     Predict           209.5360
 Mean Pct Error      Use                -7.4454
                     Predict            -3.2476
 Mean Abs Pct Err    Use                20.4502
                     Predict             8.7346
 SSE                 Use            43096722.4
                     Predict        438747.710
 MSE                 Use            552522.082
                     Predict        54843.4638
 RMS                 Use              743.3183
                     Predict          234.1868
 Durbin-Watson       Use                1.8051
                     Predict            1.4246
```

Como se puede observar, la mayoría de los valores de los estadísticos obtenidos, en el período de validación son consecuentes e incluso mejores que los del período de estimación (excepto Durbin-Watson) pero no se debe estar conforme con este modelo entre

otras cosas porque el resultado de Durbin-Watson puede mejorarse, entiéndase por mejorar, lograr que este estadístico se acerque lo más posible al valor 2. Si se regresa a la idea original que la serie no era estacionaria regularmente y además a la idea de diferenciar estacionalmente, se obtiene en la práctica un modelo mucho mejor que este y que responde a la forma (0 1 1)(0 1 1)4:

```
ARIMA place..e
 /MODEL=( 0 1 1 )( 0 1 1 ) NOCONSTANT
 /MXITER 10
 /PAREPS .001
 /SSQPCT .001
 /FORECAST EXACT .
Split group number: 1  Series length: 80
FINAL PARAMETERS:
```

```
Number of residuals  75
Standard error       763.54405
Log likelihood       -604.32151
AIC                  1212.643
SBC                  1217.278

Analysis of Variance:
              DF  Adj. Sum of Squares    Residual Variance
Residuals     73           43787119.1            582999.52
           Variables in the Model:
                B          SEB     T-RATIO   APPROX. PROB.
MA1     .83825228   .07347532   11.408624       .00000000
SMA1    .35422532   .12160482    2.912922       .00474839
```

Chequeo diagnóstico de los residuales:

```
Autocorrelations:   ERR_3   Error for PLACE..E from ARIMA, MOD_21 NO
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                  ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  1  -.044   .113                .   *⬄    .                         .149   .700
  2  -.005   .112                 .   *   .                          .151   .927
  3   .004   .112                 .   *   .                          .152   .985
  4   .036   .111                 .   ⬄*  .                          .260   .992
  5   .188   .110                 .   ⬄****                         3.162   .675
  6  -.065   .109                 .  *⬄   .                         3.521   .741
  7   .011   .109                 .   *   .                        3.531   .832
  8  -.137   .108                 .***⬄   .                         5.154   .741
  9  -.051   .107                 .  *⬄   .                         5.382   .800
 10  -.037   .106                 .  *⬄   .                         5.502   .855
 11  -.041   .105                 .  *⬄   .                         5.657   .895
 12   .117   .104                 .   ⬄** .                         6.905   .864
 13  -.127   .104                 .***⬄   .                         8.400   .817
 14   .023   .103                 .   *   .                        8.450   .865
 15   .185   .102                 .   ⬄****                        11.753   .698
 16  -.156   .101                 .***⬄   .                        14.148   .588
Partial Autocorrelations:   ERR_3   Error for PLACE..E from ARIMA, MOD_21 NO
 Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                  ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  1  -.044   .115                .   *⬄    .
  2  -.007   .115                .    *    .
  3   .003   .115                .    *    .
  4   .037   .115                .    ⬄*    .
  5   .191   .115                .    ⬄****.
  6  -.049   .115                .   *⬄    .
  7   .008   .115                .    *    .
  8  -.147   .115                . ***⬄    .
  9  -.082   .115                .  **⬄    .
 10  -.082   .115                .  **⬄    .
 11  -.026   .115                .   *⬄    .
```

```
12   .132   .115                .    ⇔*** .
13  -.054   .115                .   *⇔    .
14   .040   .115                .    ⇔*   .
15   .219   .115                .    ⇔****.
16  -.185   .115                .****⇔    .
```

Un paso complementario en el diagnóstico, es comprobar hasta que punto la serie estimada es capaz de pronosticar bien los valores reservados para la validación, **ver anexo III-5.**

Ello corrobora que se está en presencia de un buen modelo; note que todas las autocorrelaciones quedan dentro de la banda de confianza alrededor de cero, y por tanto no resultan significativas. El test de Box-Ljung en cada valor de h informa que no hay razones para sospechar que dichas autocorrelaciones difieren de las correspondientes a un ruido blanco.

Ahora, para concluir existen principalmente dos variantes a la hora de elegir cual modelo es mejor, una es analizando los siguientes criterios.

Modelos	Estándar Error	Log likelihood	AIC	SBS
(0 0 0)(1 0 0)4	681.23523	-699.28884	1402.5777	1407.5324
(0 1 1)(0 1 1)4	731.79717	-665.37668	1334.7534	1339.591

Como se puede apreciar, para el modelo (0 1 1)(0 1 1)4 los criterios de AIC y SBS son mejores, (es decir, arrojan valores menores y por tanto reducción de costo respecto a beneficios) y según este criterio se puede concluir que el mejor modelo es el último. La segunda variante es utilizar el comando FIT, con el objetivo de obtener algunos estadísticos interesantes de diagnóstico y que son especialmente útiles para comparar los modelos estudiados.

```
                        FIT Error Statistics
Error Variable                       ERR_2       ERR_4
Observed Variable                    PLACE..E    PLACE..E
N of Cases          Use               80          75
                    Predict            8           8
Deg Freedom         Use               78          73
                    Predict            8           8
Mean Error          Use                2.7102      9.0222
                    Predict          -71.5321    128.8192
Mean Abs Error      Use              492.3472    524.7007
                    Predict          209.5360    271.9092
Mean Pct Error      Use               -7.4454     -3.7368
                    Predict           -3.2476      5.2555
Mean Abs Pct Err    Use               20.4502     22.4981
                    Predict            8.7346     11.0736
SSE                 Use            43096722.4  43490477.9
                    Predict        438747.710  893688.324
MSE                 Use            552522.082  595759.971
                    Predict        54843.4638  111711.041
RMS                 Use              743.3183    771.8549
                    Predict          234.1868    334.2320
Durbin-Watson       Use                1.8051      2.0559
```

	Predict	1.4246	1.7545

Note que entre ambos modelos existen pocas diferencias entre los estadísticos MAPE, SSE, MSE, RMS, pero mejores (menores) para el primer modelo analizado. Sin embargo, teniendo en cuenta que la condición de tener los residuales incorrelacionados es esencial y que Durbin-Watson es más cercano a 2 en el segundo modelo, podemos llegar a la conclusión que el modelo más eficiente obtenido entonces para la serie consumo de energía eléctrica en el municipio de Placetas responde a la forma (0 1 1)(0 1 1)4:

$(1-B)(1-B^S)X_t = (1-MA1B)(1-SMA1B^S)e_t$ o

$(1-B)(1-B^4)X_t = (1-0.8485B)(1-0.3889B^4)e_t$ o también más concretamente:

$$X_t=X_{t-1}+X_{t-4} - X_{t-5}+e_t -0.8485e_{t-1}-0.3889\ e_{t-4}+0.32998\ e_{t-5}$$

Donde $\mathbf{e_t}$ es un ruido blanco $(0,\sigma^2)$ con σ^2 =535527.10

Se puede resumir el diagnóstico del modelo de la siguiente forma:

Sobre	Resultados
El proceso	Se alcanza la convergencia y en particular se mantiene controlada la constante de Marquart.
Los parámetros	Resultan todos significativos
Correlación entre parámetros	No existe
Redundancia AR-MA	No es analizable en este caso
Media de los residuales	No cercana a cero
Varianza de los residuales	Gráficamente parece constante.
Correlación de los residuales	No significativos
Ajuste a ruido blanco	No se rechaza por Box-Ljung
Validación	Pronostica bien los datos reservados
Estadísticos de error	Se comportan estables en la fase de cálculo y de pronóstico.

Para concluir, se muestra en el **anexo III-6**, aunque no fue la seleccionada, tanto la serie original como la serie Fit_2 del modelo (0 0 0)(1 0 0)4 así como la serie original como la serie Fit_4 para el modelo seleccionado (0 1 1)(0 1 1)4 para que se pueda tener una idea del grado de concordancia entre ambas series.

3.2 Serie consumo eléctrico de Santa Clara con modelo ARIMA

En este epígrafe se tratará la serie consumo eléctrico de Santa Clara, pero a diferencia del capítulo 2, en el cual se trabajó esta misma serie con el enfoque clásico, se tratará en esta

ocasión con modelos ARIMA, utilizando la metodología Box-Jenkins, para luego realizar una comparación.

Se trabajará con la serie más estable es decir la serie a la cual se le sustituyeron los outliers por cierta interpolación lineal. Además, después de un intenso proceso de búsqueda se llegó a que el mejor modelo es (2 1 0)(1 1 0)4, pero antes de considerar como definitiva esta alternativa y teniendo en cuenta que gráficamente se nota un descenso a finales del año 1990 y principio del 91, que coincide además con el año que oficialmente en el país se reconoce como inicio del período especial, se introduce el regresor**" reg90Q4"** como un indicador de esta etapa. El resultado del modelo con el regresor se muestra a continuación:

```
ARIMA santa._1 WITH reg90q4
 /MODEL=( 2 1 0 )( 1 1 0 ) NOCONSTANT
 /MXITER 10
 /PAREPS .001
 /SSQPCT .001
 /FORECAST EXACT .
```

```
Split group number: 1  Series length: 88
FINAL PARAMETERS:
Number of residuals  83
Standard error       4321,5842
Log likelihood       -811,01899
AIC                  1630,038
SBC                  1639,7133
            Analysis of Variance:
              DF  Adj. Sum of Squares    Residual Variance
Residuals     79         1491202865,7           18676090,3
           Variables in the Model:
                     B           SEB      T-RATIO   APPROX. PROB.
AR1             -,40080       ,10920   -3,6704961       ,00043811
AR2             -,22227       ,11034   -2,0144436       ,04736611
SAR1            -,38867       ,10358   -3,7524108       ,00033268
REG90Q4    -11510,71913   3661,77923   -3,1434771       ,00235183
```

Note que todos los parámetros resultan significativos y el comportamiento de los residuales aparece como:

```
ACF
 VARIABLES= err_1
 /NOLOG
 /MXAUTO 16
 /SERROR=IND
 /PACF.
```

```
Autocorrelations:   ERR_1   Error for SANTA._1 from ARIMA, MOD_1 NOC
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                 ▫⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩▫
  1   ,000   ,108                  .   *   .                          ,000   ,997
  2  -,024   ,107                  .   *   .                          ,052   ,975
  3   ,007   ,106                  .   *   .                          ,056   ,997
  4  -,079   ,106                  . **⇔   .                           ,608   ,962
  5   ,194   ,105                  .   ⇔****                          4,019   ,547
  6   ,071   ,104                  .   ⇔*  .                          4,487   ,611
  7   ,041   ,104                  .   ⇔*  .                          4,645   ,703
  8  -,195   ,103                 ****⇔   .                           8,239   ,411
  9  -,024   ,102                  .   *   .                         8,292   ,505
 10   ,169   ,102                  .   ⇔***.                         11,058   ,353
 11  -,002   ,101                  .   *   .                        11,059   ,438
 12  -,089   ,100                  . **⇔   .                         11,849   ,458
 13  -,023   ,100                  .   *   .                        11,903   ,536
 14  -,055   ,099                  .  *⇔   .                         12,212   ,589
```

```
 15   ,089   ,098                 .   ⇔** .                    13,029   ,600
 16   ,075   ,097                 .   ⇔** .                    13,624   ,627
Partial Autocorrelations:   ERR_1   Error for SANTA._1 from ARIMA, MOD_1 NOC
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
  1   ,000   ,110                 .   *   .
  2  -,024   ,110                 .   *   .
  3   ,007   ,110                 .   *   .
  4  -,079   ,110                 . **⇔    .
  5   ,196   ,110                 .   ⇔****
  6   ,067   ,110                 .   ⇔*   .
  7   ,054   ,110                 .   ⇔*   .
  8  -,213   ,110                 ****⇔    .
  9   ,011   ,110                 .   *   .
 10   ,144   ,110                 .   ⇔***.
 11  -,012   ,110                 .   *   .
 12  -,164   ,110                 .***⇔    .
 13   ,043   ,110                 .   ⇔*   .
 14   ,012   ,110                 .   *   .
 15   ,068   ,110                 .   ⇔*   .
 16  -,021   ,110                 .   *   .
```

Los residuales no muestran ningún desajuste significativo a las de un ruido blanco. Por tanto, hemos llegado a una buena modelación.

Por otra parte, esta serie ya fue estudiada según el enfoque clásico, y los pasos que se siguieron allí (ahora las variables cambian de nombre) fueron:

```
SEASON
 /VARIABLES=santa._1
 /MODEL=ADDITIVE
 /MA=EQUAL.
```

```
Results of SEASON procedure for variable SANTA._1
Additive Model.  Equal weighted MA method.  Period = 4.
                 Seasonal
      Period        index
         1       -2318.30
         2        790.104
         3        250.108
         4       1278.088
```

Luego, se realiza la estimación cuadrática con el comando CURVEFIT resultando:

```
CURVEFIT /VARIABLES=sas_2
 /CONSTANT
 /MODEL=QUADRATIC
 /PRINT ANOVA
 /PLOT FIT
 /SAVE=PRED RESID .
```

```
Dependent variable.. SAS_2             Method.. QUADRATI
Multiple R           .85537
R Square             .73165
Adjusted R Square    .72534
Standard Error   7525.96784
            Analysis of Variance:
               DF   Sum of Squares      Mean Square
Regression      2    13126705194.0     6563352597.0
Residuals      85     4814416317.8       56640192.0
F =     115.87801       Signif F =  .0000

-------------------- Variables in the Equation --------------------

Variable                  B         SE B       Beta         T  Sig T
Time           -1264.371991  127.710378  -2.249342    -9.900  .0000
```

Time**2	9.361142	1.390376	1.529690	6.733	.0000
(Constant)	73093.772174	2462.568089		29.682	.0000

A continuación, se calculan la serie pronóstico según el modelo aditivo $\mathbf{Y_t=T_t+S_t+E_t}$ **(la cual cumple como se mostró, que E_t tiene estructura ARMA)** y la serie de residuales $\mathbf{E_t}$ mediante los comandos:

```
COMPUTE pronos.a = saf_2+fit_3.
EXECUTE.
COMPUTE error.a = santa._1-pronos.a.
EXECUTE.
```

Por último, se realiza la comparación, (objetivo central de este epígrafe) entre los modelos hallados según el enfoque clásico y según ARIMA resultando:

```
FIT
ERRORS=err_1 error.a /
OBS= FIT_1 pronos.a /
DFE=  79 85.
```

FIT Error Statistics

Error Variable		ERR_1	ERROR.A
Observed Variable		FIT_1	PRONOS.A
N of Cases	Use	83	88
Deg Freedom	Use	79	85
Mean Error	Use	121,1530	,0000
Mean Abs Error	Use	3074,0687	5769,9551
Mean Pct Error	Use	2,4152	-,0027
Mean Abs Pct Err	Use	9,5780	13,1901
SSE	Use	1477667588	4814416318
MSE	Use	18704653,0	56640192,0
RMS	Use	4324,8876	7525,9678
Durbin-Watson	Use	1,9860	,2777

Note que la mayoría de los estadísticos son mejores (menores) para la modelación según la modelación ARIMA que utilizando el enfoque clásico y teniendo en cuenta que la condición de tener los residuales incorrrelacionados es esencial, y que Durbin-Watson esta bastante cercano a 2, el modelo ARIMA es el seleccionado.

Como punto esencial en este trabajo consiste en realizar los pronósticos para el año 2006, se utiliza el modelo hallado para ejemplificar cómo se calculan los mismos. Ello es posible con:

```
ARIMA santa._1 WITH reg90q4
  /MODEL=( 2 1 0 )( 1 1 0 ) NOCONSTANT
  /MXITER 10
  /PAREPS .001
  /SSQPCT .001
  /FORECAST EXACT .
```

resultando:

DATE	Fit	Lcl	ucl	sep	Valores reales
Q1 2006	27193,48326	18591,58500	35795,38151	4321,58424	27791.00
Q2 2006	31731,13277	23129,23452	40333,03102	4321,58424	35882.60
Q3 2006	31894,31368	23292,41543	40496,21194	4321,58424	33897.20
Q4 2006	35345,96883	26744,07058	43947,86709	4321,58424	31129.40

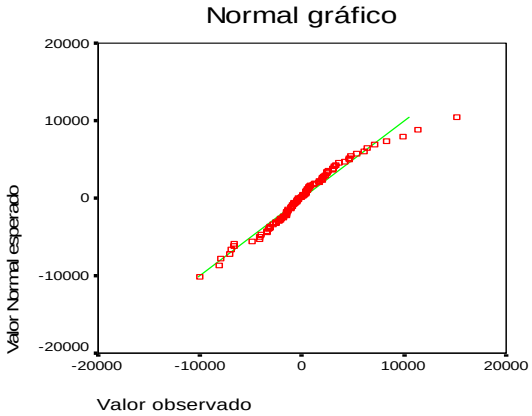
Note que la menor diferencia es de 597 Mwh en el primer trimestre y respecto a los demás trimestres los estimados no son tan buenos pero se pueden considerar como aceptables hasta cierto punto si se tiene en cuenta la complejidad de la serie analizada aunque esta situación debe mejorarse con la inclusión de nuevos datos provenientes de los años reales 2007 y 2008, y/o con el hallazgo de un modelo más eficiente.

Por último, en ésta como en todas las series del presente trabajo, interesan especialmente los intervalos de confianza del pronóstico, por ello se verifica además si los residuales se distribuyen normalmente.

Para ello históricamente se aplica la prueba de Kolmogorov-Smirnov pero esta prueba en nuestro caso puede ser cuestionada por razones de potencia ya que es aplicable solamente si los datos sobrepasan los 100 valores y este no es el caso, pero el módulo TRENDS del SPSS/PC brinda al menos la posibilidad de un análisis gráfico de la normalidad de series (y en particular residuales) mediante el comando NPPLOT, que contrasta valores observados y esperados bajo supuesto de distribución normal. Todo ello se logra con el comando:

```
NPPLOT err_1.
```

y el resultado es:



Como no hay una divergencia aparente significativa de la recta y=x (frecuencias esperadas = frecuencias observadas), podemos admitir que nuestros residuales se distribuyen normalmente.

Por otra parte podemos contar también con un test de Normalidad con la corrección de la significación de Lilliefors, la cual es aplicable cuando la cantidad de datos analizados se encuentran entre 50 y 100 respectivamente lo cual es perfectamente aplicable en nuestro caso, lográndose con el siguiente comando.

```
EXAMINE
 VARIABLES=err_2
 /PLOT BOXPLOT STEMLEAF NPPLOT
 /COMPARE GROUP
 /STATISTICS NONE
```

```
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

Pruebas de normalidad

	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
Error for SANTA._1 from ARIMA, MOD_1 NOCON	,081	83	,200*

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

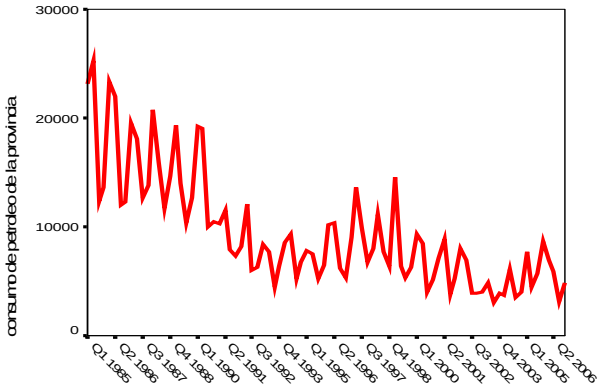
Este comando se activa con las opciones **Analizar \ Estadísticos descriptivos \ Explorar \Gráficos**. Se trata de la prueba de Kolmogorov-Smirnov con la corrección de Lilliefors.

que permite realizar pruebas de normalidad en muestras de tamaño entre 50 y 100 respectivamente. Observe el valor de significación 0.200 con lo cual se corrobora el análisis gráfico por lo que podemos decir que no se rechaza la normalidad para la muestra.

3.3. Serie Consumo Provincial de Petróleo (Ton)

Lo primero que salta a la vista cuando se observa el gráfico de esta serie, es el considerable descenso, que coincide con el inicio del período especial en el año 1991, lo cual se muestra claramente en el rango de estos valores que es de 22125 toneladas de petróleo.

	N	Rango	Mínimo	Máximo	Media
consumo de Petróleo de la provincia	88	22125.80	3092.10	25217.90	9465.7295



De este ploteo resulta evidente una tendencia lineal decreciente, una posible tendencia periódica, así como una aparente homocedasticidad. Corresponde según la metodología,

hacer el cálculo y ploteo de las funciones de autocorrelación y autocorrelación parcial, lo cual se realiza sólo con los primeros 80 casos:

ACF
VARIABLES= petroleo
/NOLOG
/MXAUTO 16
/SERROR=IND
/PACF.

```
Autocorrelations:   PETROLEO   consumo de petroleo de la provincia
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                   ▫⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩▫
  1   ,679   ,110                .   ⇔***.**********            38,318   ,000
  2   ,415   ,109                .   ⇔***.****                  52,819   ,000
  3   ,596   ,108                .   ⇔***.********              83,072   ,000
  4   ,766   ,108                .   ⇔***.***********          133,735   ,000
  5   ,538   ,107                .   ⇔***.*******              159,012   ,000
  6   ,332   ,106                .   ⇔***.***                  168,768   ,000
  7   ,467   ,105                .   ⇔***.*****                188,344   ,000
  8   ,573   ,105                .   ⇔***.*******              218,240   ,000
  9   ,383   ,104                .   ⇔***.****                 231,773   ,000
 10   ,222   ,103                .   ⇔****                     236,393   ,000
 11   ,352   ,103                .   ⇔***.***                  248,184   ,000
 12   ,430   ,102                .   ⇔***.*****                266,000   ,000
 13   ,231   ,101                .   ⇔***.*                    271,246   ,000
 14   ,098   ,100                .   ⇔** .                     272,204   ,000
 15   ,219   ,100                .   ⇔****                     277,025   ,000
 16   ,297   ,099                .   ⇔***.**                   286,078   ,000
Plot Symbols:      Autocorrelations *     Two Standard Error Limits .
Total cases:  80     Computable first lags:  79
Partial Autocorrelations:   PETROLEO   consumo de petroleo de la provincia
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                   ▫⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩⇳⇩⇩⇩⇩▫
  1   ,679   ,112                .   ⇔***.**********
  2  -,086   ,112                . **⇔    .
  3   ,651   ,112                .   ⇔***.*********
  4   ,280   ,112                .   ⇔***.**
  5  -,197   ,112                ****⇔    .
  6  -,070   ,112                .  *⇔    .
  7   ,030   ,112                .   ⇔*   .
  8  -,038   ,112                .  *⇔    .
  9  -,065   ,112                .  *⇔    .
 10  -,047   ,112                .  *⇔    .
 11   ,094   ,112                .   ⇔**  .
 12  -,029   ,112                .  *⇔    .
 13  -,119   ,112                . **⇔    .
 14  -,033   ,112                .  *⇔    .
 15  -,024   ,112                .   *   .
 16   ,064   ,112                .   ⇔*   .
```

Los correlogramas indican que la serie no es estacionaria e indican la necesidad de emprender una posible diferenciación al menos regular; por tanto se recalculan los correlogramas de la serie diferenciada pero aumentando el número de lags para ratificar la presencia de tendencias estacionales, independientes de la tendencia lineal**, ver anexo III-7**.

Ahora resulta evidente la tendencia periódica anual de la serie y por tanto la estacionalidad S=4; ello se muestra en la ACF con la cantidad de espigas que regularmente y alternativamente salen fuera del intervalo de confianza. Con el objetivo de identificar la estructura estacional (P, D, Q) del modelo, se trazan los autocorrelogramas estacionales:

ACF

```
VARIABLES= petroleo
/NOLOG
/DIFF=1
/MXAUTO 60
/SERROR=IND
/SEASONAL
/PACF.
```

```
Autocorrelations:   PETROLEO   consumo de petroleo de la provincia
Transformations:  difference (1)
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                  ▫⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩▫
  4   ,622   ,108                  .   ⇔***.********              71,133   ,000
  8   ,460   ,105                  .   ⇔***.*****                113,199   ,000
 12   ,413   ,102                  .   ⇔***.****                 150,099   ,000
 16   ,375   ,099                  .   ⇔***.***                  180,493   ,000
 20   ,334   ,096                  .   ⇔***.***                  210,771   ,000
 24   ,353   ,093                  .   ⇔***.***                  238,283   ,000
 28   ,306   ,089                  .   ⇔***.**                   263,256   ,000
 32   ,225   ,086                   .  ⇔**.**                    278,988   ,000
 36   ,241   ,082                   .  ⇔**.**                    296,820   ,000
 40   ,289   ,078                   .  ⇔**.***                   317,983   ,000
 44   ,304   ,074                   .  ⇔**.***                   349,711   ,000
 48   ,262   ,070                   .  ⇔**.**                    378,550   ,000
 52   ,236   ,065                   .  ⇔**.**                    402,982   ,000
 56   ,144   ,060                    . ⇔*.*                      424,109   ,000
 60   ,135   ,054                    . ⇔*.*                      438,593   ,000
Partial Autocorrelations:   PETROLEO   consumo de petroleo de la provincia
Transformations:  difference (1)
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                  ▫⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩▫
  4   ,037   ,113                  .    ⇔*   .
  8   ,114   ,113                  .    ⇔**  .
 12   ,040   ,113                  .    ⇔*   .
 16  -,117   ,113                  .  **⇔    .
 20  -,063   ,113                  .   *⇔    .
 24   ,106   ,113                  .    ⇔**  .
 28  -,082   ,113                  .  **⇔    .
 32   ,013   ,113                  .    *    .
 36   ,054   ,113                  .    ⇔*   .
 40   ,089   ,113                  .    ⇔**  .
 44  -,019   ,113                  .    *    .
 48   ,042   ,113                  .    ⇔*   .
 52   ,009   ,113                  .    *    .
 56  -,139   ,113                  . ***⇔    .
 60   ,073   ,113                  .    ⇔*   .
Total cases:  80     Computable first lags after differencing:  78
```

Del análisis de estos correlogramas estacionales se comprende que la serie requiera tal vez una diferenciación estacional para luego identificar los órdenes P y Q lo cual se logra con el comando:

```
ACF
 VARIABLES= petroleo
 /NOLOG
 /DIFF=1
 /SDIFF=1
 /MXAUTO 60
 /SERROR=IND
 /SEASONAL
 /PACF.
```

```
Autocorrelations:   PETROLEO   consumo de petroleo de la provincia
Transformations:  difference (1), seasonal difference (1 at 4)
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                  ▫⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩▫
  4  -,292   ,111                **.***⇔   .                      16,192   ,003
```

```
  8  -,122   ,108                 . **⬄    .                     20,625   ,008
 12   ,034   ,104                 .   ⬄*   .                     22,437   ,033
 16   ,001   ,101                 .    *   .                    28,347   ,029
 20  -,088   ,098                 . **⬄    .                     31,819   ,045
 24   ,162   ,094                 .   ⬄***.                      35,456   ,062
 28  -,020   ,090                 .    *   .                    37,460   ,109
 32  -,101   ,086                  .**⬄   .                      43,279   ,088
 36  -,057   ,082                  . *⬄   .                      51,095   ,049
 40   ,048   ,078                  .  ⬄* .                       58,785   ,028
Partial Autocorrelations:   PETROLEO   consumo de petroleo de la provincia
Transformations:  difference (1), seasonal difference (1 at 4)
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
 
  4  -,427   ,115             ****.****⬄     .
  8  -,275   ,115                 *****⬄     .
 12   ,035   ,115                .     ⬄*    .
 16  -,088   ,115                .  **⬄      .
 20  -,026   ,115                .   *⬄      .
 24   ,134   ,115                .     ⬄*** .
 28  -,085   ,115                .  **⬄      .
 32  -,110   ,115                .  **⬄      .
 36  -,089   ,115                .  **⬄      .
 40  -,044   ,115                .   *⬄      .
```

Note que ahora, la SACF(h) muestra una declinación a cero en forma sinusoidal y la SPACF(h) muestra dos espigas en los retardos h=1 y h=2; por tanto se elige en principio el modelo autorregresivo estacional de orden 2.

```
ARIMA petroleo
 /MODEL=( 0 1 0 )( 2 1 0 ) NOCONSTANT
 /MXITER 10
 /PAREPS .001
 /SSQPCT .001
 /FORECAST EXACT .
```

```
Split group number: 1  Series length: 80
FINAL PARAMETERS:

Number of residuals  75
Standard error       2726,8113
Log likelihood       -699,16899
AIC                  1402,338
SBC                  1406,973
            Analysis of Variance:
              DF  Adj. Sum of Squares    Residual Variance
Residuals     73          549215559,3           7435499,9
           Variables in the Model:
                 B          SEB      T-RATIO   APPROX. PROB.
SAR1    -,37615650   ,11414084   -3,2955469       ,00151807
SAR2    -,24622157   ,11388268   -2,1620634       ,03389121
```

Los dos parámetros SAR1 y SAR2 son significativos. Los residuales del estudio estacional no muestran en su ploteo tendencias aparentes (se omite este paso) y se precisa a continuación hasta que punto no están correlacionados estacionalmente, **ver anexo III-8.**

Una vez verificado que los errores no están muy correlacionados estacionalmente (el test de Box-Ljung en casi todos los valores de h es no significativo), se prosigue el estudio, y se pasa analizar si se requieren componentes regulares en el modelo. Para ello se regresa al trazado de los correlogramas regulares hasta un máximo que se puede reducir a 16.

ACF
 VARIABLES= err_1
 /NOLOG
 /MXAUTO 16
 /SERROR=IND
 /PACF.

```
Autocorrelations:   ERR_1   Error for PETROLEO from ARIMA, MOD_53 NO
     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                  ¤⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩¤
  1  -,287   ,113                *.****⇔    .                      6,443   ,011
  2  -,210   ,112                  ****⇔   .                       9,947   ,007
  3  -,027   ,112                  .  *⇔   .                      10,006   ,019
  4  -,017   ,111                  .   *   .                      10,031   ,040
  5   ,070   ,110                  .   ⇔*  .                      10,433   ,064
  6   ,017   ,109                  .   *   .                      10,456   ,107
  7  -,012   ,109                  .   *   .                      10,468   ,164
  8  -,034   ,108                  .  *⇔   .                      10,570   ,227
  9   ,032   ,107                  .   ⇔*  .                      10,662   ,300
 10  -,025   ,106                  .   *   .                      10,715   ,380
 11   ,178   ,105                  .   ⇔****                      13,562   ,258
 12  -,129   ,104                  .***⇔   .                      15,076   ,237
 13  -,182   ,104                  ****⇔   .                      18,170   ,151
 14   ,081   ,103                  .   ⇔** .                      18,784   ,173
 15   ,103   ,102                  .   ⇔** .                      19,797   ,180
 16  -,024   ,101                  .   *   .                      19,853   ,227
Partial Autocorrelations:   ERR_1   Error for PETROLEO from ARIMA, MOD_53 NO
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                  ¤⇩⇩⇩⇩⇩⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩¤
  1  -,287   ,115                *.****⇔    .
  2  -,319   ,115                *.****⇔    .
  3  -,241   ,115                 *****⇔    .
  4  -,239   ,115                 *****⇔    .
  5  -,130   ,115                 . ***⇔    .
  6  -,095   ,115                 .  **⇔    .
  7  -,063   ,115                 .   *⇔    .
  8  -,078   ,115                 .  **⇔    .
  9  -,016   ,115                 .    *   .
 10  -,052   ,115                 .   *⇔    .
 11   ,213   ,115                 .    ⇔****.
 12   ,058   ,115                 .    ⇔*   .
 13  -,125   ,115                 . ***⇔    .
 14  -,088   ,115                 .  **⇔    .
 15  -,020   ,115                 .    *   .
 16  -,083   ,115                 .  **⇔    .
Plot Symbols:      Autocorrelations *     Two Standard Error Limits .
Total cases:  80     Computable first lags:  74
```

Ante estos correlogramas, existen las alternativas de elegir como continuación del estudio un modelo de medias móviles o un modelo autorregresivo de orden 2 para la parte regular por aparecer una espiga en la ACF y dos en la PACF respectivamente, por tanto el autor se decide por el modelo (2 1 0)(2 1 0).Además teniendo en cuenta que gráficamente se nota un descenso alrededor del año 1991 que coincide además con el año que oficialmente en el país se reconoce como inicio del período especial, se introduce el regresor “reg1991” como variable independiente, que diferenciada una vez regularmente y una vez estacionalmente, conducen a una función pulso en el 1er trimestre del 1991 (epígrafe 1.9.1).

```
ARIMA petroleo WITH reg1991
 /MODEL=( 2 1 0 )( 2 1 0 ) NOCONSTANT
 /MXITER 10
 /PAREPS .001
 /SSQPCT .001
 /FORECAST EXACT .
```

```
Split group number: 1  Series length: 80
FINAL PARAMETERS:
Number of residuals  75
Standard error       2392,9732
Log likelihood       -688,07803
AIC                  1386,1561
SBC                  1397,7435
            Analysis of Variance:
              DF  Adj. Sum of Squares    Residual Variance
Residuals     70          408123527,6            5726320,9
           Variables in the Model:
                     B          SEB      T-RATIO   APPROX. PROB.
AR1            -,39595       ,11084   -3,5724053       ,00064515
AR2            -,35167       ,11300   -3,1121065       ,00268896
SAR1           -,39309       ,11708   -3,3574237       ,00127509
SAR2           -,27439       ,11553   -2,3751019       ,02028774
REG1991    -5190,27886   1906,94185   -2,7217814       ,00818734
```

Note que todos los parámetros resultan significativos y el comportamiento de los residuales parece satisfactorio:

```
ACF
 VARIABLES= err_3
 /NOLOG
 /MXAUTO 16
 /SERROR=IND
 /PACF.
```

```
Autocorrelations:   ERR_3   Error for PETROLEO from ARIMA, MOD_58 NO

     Auto- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1   Box-Ljung  Prob.
                   ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  1  -,079   ,113               .  **⬄    .                       ,487   ,485
  2  -,131   ,112                .***⬄   .                       1,851   ,396
  3  -,221   ,112                ****⬄   .                       5,785   ,123
  4  -,041   ,111                .  *⬄   .                       5,922   ,205
  5   ,154   ,110                .   ⬄***.                       7,890   ,162
  6   ,056   ,109                .   ⬄*  .                       8,152   ,227
  7   ,012   ,109                .   *   .                      8,164   ,318
  8  -,056   ,108                .  *⬄   .                       8,434   ,392
  9   ,076   ,107                .   ⬄** .                       8,945   ,442
 10   ,094   ,106                .   ⬄** .                       9,733   ,464
 11   ,121   ,105                .   ⬄** .                      11,059   ,438
 12  -,179   ,104                ****⬄   .                      13,997   ,301
 13  -,182   ,104                ****⬄   .                      17,077   ,196
 14   ,093   ,103                .   ⬄** .                      17,889   ,212
 15   ,108   ,102                .   ⬄** .                      19,022   ,213
 16  -,013   ,101                .   *   .                     19,039   ,267
Plot Symbols:      Autocorrelations *     Two Standard Error Limits .
Total cases:  80     Computable first lags:  74
Partial Autocorrelations:   ERR_3   Error for PETROLEO from ARIMA, MOD_58 NO
   Pr-Aut- Stand.
Lag  Corr.   Err. -1  -.75  -.5 -.25   0   .25  .5   .75   1
                   ▫⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩⇧⇩⇩⇩⇩▫
  1  -,079   ,115               .  **⬄    .
  2  -,138   ,115               . ***⬄    .
  3  -,251   ,115              *****⬄    .
  4  -,120   ,115               .  **⬄    .
  5   ,071   ,115               .    ⬄*   .
  6   ,008   ,115               .    *    .
  7   ,024   ,115               .    *    .
  8   ,010   ,115               .    *    .
  9   ,123   ,115               .    ⬄**  .
```

```
 10   ,132   ,115                .    ⇔*** .
 11   ,193   ,115                .    ⇔****.
 12  -,076   ,115                .  **⇔    .
 13  -,141   ,115                . ***⇔    .
 14   ,062   ,115                .    ⇔*   .
 15   ,011   ,115                .    *    .
 16  -,145   ,115                . ***⇔    .
```

Ya estamos en presencia de un buen modelo, pero podríamos preguntarnos si podemos mejorarlo, pues todavía en el retardo 3 de las autocorrelaciones estacionales existe una espiga que puede reflejar que los residuales responden todavía a un modelo (3 0 0) y por tanto sugieren que la serie original responda mejor al modelo (3 1 0)(2 1 0), el cual se ensaya con el comando siguiente

```
ARIMA petroleo WITH reg1991
 /MODEL=( 3 1 0 )( 2 1 0 ) NOCONSTANT
 /MXITER 10
 /PAREPS .001
 /SSQPCT .001
 /FORECAST EXACT .
```

```
Split group number: 1  Series length: 80
FINAL PARAMETERS:
Number of residuals  75
Standard error       2326,8243
Log likelihood       -685,47917
AIC                  1382,9583
SBC                  1396,8633

            Analysis of Variance:

              DF  Adj. Sum of Squares    Residual Variance

Residuals     69          380496250,1             5414111,3

           Variables in the Model:
                      B           SEB      T-RATIO   APPROX. PROB.

AR1             -,51095       ,11717   -4,3606612       ,00004433
AR2             -,50564       ,12662   -3,9932679       ,00016056
AR3             -,34191       ,14401   -2,3742872       ,02036986
SAR1            -,57935       ,14868   -3,8965869       ,00022307
SAR2            -,26330       ,12627   -2,0851706       ,04075541
REG1991    -3434,37701   1535,86063   -2,2361254       ,02858005
```

Note que todos los parámetros resultan significativos y el comportamiento de los residuales se puede ver en **Anexo III-9:**

Como se pudo observar todas las autocorrelaciones quedan dentro de la banda de confianza alrededor de cero, y por tanto no resultan significativas. El test de Box-Ljung en cada valor de h informa que no existen razones para sospechar que dichas autocorrelaciones difieren de las correspondientes a un ruido blanco. Tampoco aparecen autocorrelaciones parciales significativas en los residuales. En pocas palabras, los errores son no correlacionados.

Utilicemos el comando FIT, con el objetivo de obtener algunos estadísticos interesantes de diagnóstico y que son especialmente útiles para comparar los modelos estudiados:

```
FIT
/ERROR=ERR_3 ERR_4
/OBS=FIT_3 FIT_4
/DFE=70 69.
```

FIT Error Statistics

		(210)(210)reg1991	(310)(210)reg1991
Error Variable		ERR_3	ERR_4
Observed Variable		FIT_3	FIT_4
N of Cases	Use	75	75
	Predict	8	8
Deg Freedom	Use	70	69
	Predict	8	8
Mean Error	Use	214,5052	220,7429
	Predict	1546,0948	2195,2947
Mean Abs Error	Use	1812,9318	1731,2679
	Predict	1981,3323	2389,2441
Mean Pct Error	Use	10,4766	8,6998
	Predict	39,9841	63,9538
Mean Abs Pct Err	Use	24,7224	22,1063
	Predict	47,5551	67,6280
SSE	Use	408665324	381132344
	Predict	48375901,7	66115995,1
MSE	Use	5838076,05	5523657,16
	Predict	6046987,71	8264499,38
RMS	Use	2416,2111	2350,2462
	Predict	2459,0624	2874,8042
Durbin-Watson	Use	2,1065	2,0845
	Predict	1,1327	,7504

Evidentemente, los dos modelos son buenos ya que todos los parámetros y en particular los más importantes: la media de los errores en valores absolutos (MAPE), la media de los errores al cuadrado (MSE), la suma de cuadrado de los errores (SSE), están bastante cercanos, pero en todos los casos mejores para el modelo (3 1 0)(2 1 0)4 reg1991 . Además, teniendo en cuenta que el Durbin-Watson se comporta bastante cercano a 2 en la fase de uso podemos llegar a la conclusión que el modelo más eficiente para la serie consumo de petróleo de la provincia de Villa Clara responde a la forma (3 1 0)(2 1 0)4 con el regresor "**reg 1991**":

Ahora bien como parte del diagnóstico complementario, se genera el pronóstico paso a paso sobre el período de validación. Lo cual se logra mediante:

```
ARIMA petroleo WITH reg1991
 /MODEL=( 3 1 0 )( 2 1 0 ) NOCONSTANT
 /MXITER 10
 /PAREPS .001
 /SSQPCT .001
 /FORECAST EXACT .
```

```
Split group number: 1  Series length: 88
FINAL PARAMETERS:
Number of residuals  83
Standard error       2359,9003
```

```
Log likelihood       -760,1438
AIC                  1532,2876
SBC                  1546,8006
            Analysis of Variance:
              DF  Adj. Sum of Squares    Residual Variance
Residuals     77          437319271,5            5569129,6
           Variables in the Model:
                      B          SEB      T-RATIO   APPROX. PROB.
AR1            -,50411       ,11015   -4,5767673       ,00001781
AR2            -,49473       ,12022   -4,1152264       ,00009619
AR3            -,31316       ,13565   -2,3086943       ,02364458
SAR1           -,62374       ,13988   -4,4592053       ,00002761
SAR2           -,30415       ,12298   -2,4731123       ,01559775
REG1991    -3370,37009   1534,21624   -2,1968025       ,03104262
```

Los parámetros se mantienen con significación estadística y los residuales se pueden observar en el **anexo III-10**.

Por tanto, las autocorrelaciones no muestran ningún desajuste significativo a las de un ruido blanco. Al parecer se ha llegado a una buena modelación de la serie y el modelo es válido.

Aunque gráficamente en la serie original, no se puede apreciar una marcada tendencia a la recuperación en este rubro, el autor quiso considerar además el regresor "recuperación del período especial" el cual se introdujo con el nombre de **reg2000** (y que se supone actúe de forma instantánea en el año 2000) resultando:

```
ARIMA petroleo WITH reg1991 reg2000
 /MODEL=( 3 1 0 )( 2 1 0 ) NOCONSTANT
 /MXITER 10
 /PAREPS .001
 /SSQPCT .001
 /FORECAST EXACT .
```

```
Split group number: 1  Series length: 88
FINAL PARAMETERS:

Number of residuals  83
Standard error       2366,0934
Log likelihood       -759,85895
AIC                  1533,7179
SBC                  1550,6498
            Analysis of Variance:
              DF  Adj. Sum of Squares    Residual Variance
Residuals     76          433992628,0            5598398,0
           Variables in the Model:
                      B          SEB      T-RATIO   APPROX. PROB.
AR1            -,49070       ,10956   -4,4789287       ,00002602
AR2            -,49170       ,11947   -4,1157963       ,00009701
AR3            -,32324       ,13534   -2,3883601       ,01940674
SAR1           -,62614       ,13893   -4,5067041       ,00002347
SAR2           -,31132       ,12336   -2,5237606       ,01369792
REG1991    -3350,27313   1539,20297   -2,1766285       ,03261413
REG2000    -1213,39617   1539,20297    -,7883276       ,43295720
```

El reporte del SPSS indica que el mismo no es significativo y por tanto ello justifica estadísticamente el hecho de que no se puede hablar todavía de una franca recuperación en este rubro en este año.

Ahora para concluir, existen principalmente dos variantes a la hora de elegir cual modelo es mejor una es analizando los siguientes criterios.

Modelos	Estándar Error	Log likelihood	AIC	SBS
(3 1 0)(2 1 0)4 con reg 1991	**2359.9003**	**-760.143847917**	**1532.2876**	**1546.8006**
(3 1 0)(2 1 0)4 con reg 1991y reg 2000	**2366.0934**	**-759.85895**	**1533.7179**	**1550.6498**

Como se puede apreciar para el modelo (3 1 0)(2 1 0)4 con **reg 1991** los criterios de AIC y SBS son mejores es decir menores y es por eso que se puede decidir que el mejor modelo es el seleccionado. La segunda variante de comparación es utilizando el comando FIT, el cual se logra con el comando:

```
FIT
/ERROR=ERR_5 ERR_6
/OBS=FIT_5 FIT_6
/DFE=69 76
```

FIT Error Statistics

(310)(210)reg1991 (310)(210)reg1991y reg2000

Error Variable		ERR_5	ERR_6
Observed Variable		FIT_5	FIT_6
N of Cases	Use	75	75
	Predict	8	8
Deg Freedom	Use	75	75
	Predict	8	8
Mean Error	Use	223,2721	293,8512
	Predict	113,2123	124,8711
Mean Abs Error	Use	1742,1855	1720,6690
	Predict	2360,7145	2377,7195
Mean Pct Error	Use	8,9178	10,1026
	Predict	15,9543	16,5370
Mean Abs Pct Err	Use	22,4670	22,5840
	Predict	48,8370	49,4753
SSE	Use	382825258	379323255
	Predict	54780053,3	55032095,4
MSE	Use	5104336,78	5057643,40
	Predict	6847506,66	6879011,93
RMS	Use	2259,2779	2248,9205
	Predict	2616,7741	2622,7871
Durbin-Watson	Use	2,0792	2,0815
	Predict	1,8425	1,8723

Evidentemente, los dos modelos son buenos ya que todos los parámetros y en particular los más importantes: la media de los errores en valores absolutos (MAPE), los errores medios (Mean Error) están bastante cercanos pero mejores para el modelo (3 1 0)(2 1 0) con **reg 1991**.Además, teniendo en cuenta que el regresor "**reg 2000**" no fue significativo estadísticamente y que el Durbin-Watson se comporta bastante cercano a 2 en la fase de uso podemos llegar a la conclusión que el modelo más eficiente obtenido entonces para la

serie consumo de petróleo de la provincia de Villa Clara responde a la forma (3 1 0)(2 1 0)4 con el regresor **reg 1991**.

$$(1-AR1\ B-AR2\ B^2-AR3B^3)(1-SAR1B^4-SAR2\ B^8)(1-B)(1-B^4)X_t=e_t-(3370.37009)\bullet reg1991$$

o también más concretamente:

$$X_t= 0.49589\ X_{t-1} + 0.00938\ X_{t-2} + 0.18157\ X_{t-3} + 0.68942\ X_{t-4} - 0.186584\ X_{t-5} - 0.00352932\ X_{t-6} - 0.0683175\ X_{t-7} + 0.20176\ X_{t-8} - 0.158481\ X_{t-9} - 0.00299775\ X_{t-10} - 0.058028\ X_{t-11} + 0.204067\ X_{t-12} - 0.150825\ X_{t-13} - 0.08285293\ X_{t-14} - 0.00552245\ X_{t-15} - 0.0952476\ X_{t-16} + e_t - (3370.37009)\bullet reg1991$$

donde $\mathbf{e_t}$ es un ruido blanco y $\sigma^2 \approx$ 5569129.

Una idea gráfica del buen grado de coincidencia de la serie original con la serie estimada Fit_5 se puede ver en el **anexo III-11.** Por último, en ésta como en todas las series del presente trabajo, interesan especialmente los intervalos de confianza del pronóstico, por ello se verifica además si los residuales se distribuyen normalmente.

Como habíamos mencionado anteriormente podemos contar también con un test de Normalidad con la corrección de la significación de lilliefors, la cual es aplicable cuando la cantidad de datos analizados se encuentran entre 50 y 100 respectivamente lo cual es perfectamente aplicable en nuestro caso, **ver anexo III-12.**

Observe el valor de significación 0.200 con lo cual se corrobora el análisis gráfico por lo que podemos decir que no se rechaza la normalidad para la muestra.

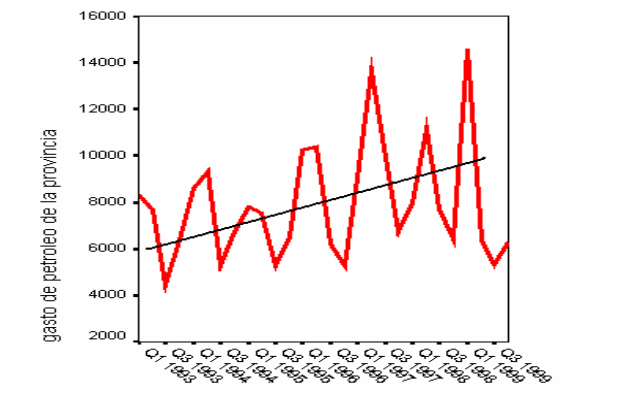
3.3.1 Análisis de la recuperación del período especial.

Este epígrafe responderá dos preguntas interesantes relacionadas con la modelación de la influencia del período especial.

1.- ¿Permite el modelo acotar la influencia del Periodo Especial?

2.- ¿Puede decidirse, a partir del modelo, si el proceso de recuperación de esta influencia negativa es abrupto o paulatino?

Como se puede ver gráficamente a partir del año 1993 hay una tendencia en la media creciente por lo que implica que hubo un aumento del consumo de petróleo en la provincia.



Utilicemos ahora un regresor que denominaremos "**RECUPERA**" para tratar de medir la influencia de esta tendencia durante el 93-99 sobre el conjunto general de la serie. Un tal regresor se implementa de manera que sea igual a 1 entre estos años y 0 fuera de los mismos

```
ARIMA petroleo WITH recupera
  /MODEL=( 2 1 0 )( 0 0 0 ) NOCONSTANT
  /MXITER= 10
  /PAREPS= .001
  /SSQPCT= .001
  /FORECAST= EXACT .
```

```
Variable:   PETROLEO
Regressors: REC9399

FINAL PARAMETERS:
Number of residuals  27
Standard error       2741,7552
Log likelihood       -250,94039
AIC                  507,88079
SBC                  511,7683
           Analysis of Variance:
             DF  Adj. Sum of Squares    Residual Variance
Residuals    24          186239285,4             7517221,5
          Variables in the Model:
                    B         SEB      T-RATIO   APPROX. PROB.
AR1          -,356144      ,16065   -2,2169357       ,03635274
AR2          -,575832      ,15936   -3,6133239       ,00139072
RECUPERA    -36,234296   281,36897    -,1287786       ,89860597
```

Como se puede observar lo parámetros **AR1** y **AR2** son significativos mientras que el regresor **RECUPERA** no lo es, por lo que no tenemos razones para suponer que haya una recuperación período 1993-1999 que marque a toda la serie. Los residuales no muestran ningún desajuste significativo a las de un ruido blanco, tal como se muestra en el **anexo III-13.**

Al ser modelable Arima con una diferenciación regular y no mostrar ningún desajuste significativo en el trazado de los correlogramas podemos decir que la serie es estacionaria y más que eso, que sigue una tendencia lineal en la media de la serie y es por eso que se

puede llegar a la conclusión que el proceso de recuperación de la influencia del período especial es de manera progresiva y paulatina.

3.4. Generalización a los restantes municipios y la provincia

Los modelos discretos hallados anteriormente, fueron explicados en detalle para ilustrar el uso de la metodología, pero el objetivo que este trabajo pretende es obtener los modelos para los cuatro municipios de mayor peso económico y la provincia los cuales resultaron ser:

MODELOS MATEMATICOS TIPO ARIMA.

SERIE CONSUMO DE ENERGÍA ELÉCTRICA DEL MUNICIPIO PLACETAS.

MODELO (0 1 1)(0 1 1)4

	B	SEB	T-RATIO	APPROX. PROB.
MA1	.84854291	.06662387	12.736320	.00000000
SMA1	.38898553	.11101820	3.503800	.00074993

$$(1-B)(1-B^4)X_t = (1-MA1B)(1- SMA1\ B^4)e_t$$

SERIE CONSUMO DE ENERGÍA ELÉCTRICA DEL MUNICIPIO CAIBARIEN.

MODELO (0 1 1)(0 1 1)4.

	B	SEB	T-RATIO	APPROX. PROB.
MA1	.72860163	.07686483	9.4789986	.0000000
SMA1	.92494477	.10112571	9.1464844	.0000000

$$(1-B)(1-B^4)X_t = (1-MA1B)(1- SMA1\ B^4)e_t$$

SERIE CONSUMO DE ENERGÍA ELÉCTRICA DEL MUNICIPIO SAGUA.

MODELO (0 1 1)(0 1 1)4.

	B	SEB	T-RATIO	APPROX. PROB.
MA1	,81114567	,07292595	11,122868	,0000000
SMA1	,90678924	,09102671	9,961793	,0000000

$$(1-B)(1-B^4)X_t = (1-MA1B)(1- SMA1\ B^4)e_t$$

SERIE CONSUMO DE ENERGÍA ELÉCTRICA DEL MUNICIPIO SANTA CLARA.

MODELO (2 1 0)(1 1 0)4 CON REGRESOR.

	B	SEB	T-RATIO	APPROX. PROB.
AR1	-,40080	,10920	-3,6704961	,00043811
AR2	-,22227	,11034	-2,0144436	,04736611
SAR1	-,38867	,10358	-3,7524108	,00033268
REG90Q4	-11510,71913	3661,77923	-3,1434771	,00235183

$$(1-B)(1-B^4)X_t = (1-MA1B)(1- SMA1\ B^4)e_t - (11510{,}71913)\cdot reg90Q4$$

SERIE CONSUMO DE PETRÓLEO DE LA PROVINICA DE VILLA CLARA

MODELO (3 1 0)(2 1 0)4 CON REGRESOR.

	B	SEB	T-RATIO	APPROX. PROB.
AR1	-,50411	,11015	-4,5767673	,00001781
AR2	-,49473	,12022	-4,1152264	,00009619
AR3	-,31316	,13565	-2,3086943	,02364458
SAR1	-,62374	,13988	-4,4592053	,00002761
SAR2	-,30415	,12298	-2,4731123	,01559775
REG1991	-3370,37009	1534,21624	-2,1968025	,03104262

$$(1-AR1\ B-AR2\ B^2-AR3B^3)(1-SAR1B^4-SAR2\ B^8)(1-B)(1-B^4)X_t=e_t-(3370.37009)\bullet reg1991$$

SERIE CONSUMO DE ENERGÍA ELÉCTRICA DE LA PROVINICA DE VILLA CLARA.
MODELO ((5) 1 1)(1 1 0)4 CON REGRESOR.

	B	SEB	T-RATIO	APPROX. PROB.
AR5	,33901	,10505	3,2270624	,00182136
MA1	,58929	,09498	6,2041803	,00000002
SAR1	-,43825	,10271	-4,2668969	,00005458
REG1991	-14823,22174	5394,96341	-2,7476038	,00743428

$$(1\text{-}AR5\ B^{5})(1\text{-}SAR1B^{4})(1\text{-}B)(1\text{-}B^{4})X_t=(1\text{-}MA1B)e_t - (14823.22174)\ reg1991$$

Por otra parte, si se realiza una revisión general de los cuatro municipios y la provincia en cuanto al consumo de energía eléctrica en el período que se está estudiando, de 1985 al 2006, mediante una tabla como la siguiente, calculada con ayuda del SPSS, y teniendo en cuenta los modelos expuestos anteriormente se pueden extraer las siguientes conclusiones:

municipios	N	Range	Minim	Maxim	Sum	mean
Santa Clara	88	47373.2	24273.1	71646.3	3694934	41987.88636364
Sagua	88	20221.1	1110.6	21331.7	979901.1	11135.23977273
Caibarién	88	7647	542.3	8189.3	174690.76	1985.122272727
Placetas	88	4277	1205.9	5482.9	233263.1	2650.717045455
PROVINCIA	N	Range	Minim	Maxim	Sum	mean
petróleo	88	22125.8	3092.1	25217.9	832984.2	9465.729545455
electricidad	88	71093.5	48883.5	119977	6726450.1	76436.93295455

1) Los municipios de Santa Clara y Sagua son los de mayor consumo en la provincia muy por encima de Placetas y Caibarién respectivamente; ello era de esperar si se tiene en cuenta la cantidad de industrias que residen en estos municipios.
2) El modelo que aparece con más frecuencia es (0 1 1)(0 1 1)4 el cual se presentó en los municipios de Placetas y Caibarién respectivamente.
3) El regresor "incidencia del período especial" que se nombró "Reg90Q4" y "Reg1991" (inicio del período especial a finales del año 1990 y principio del 1991 respectivamente) se presentó con significación estadística en las series:

Energía eléct. municipio de Santa Clara	(2 1 0)(1 1 0)4	Reg90Q4= *-11510.71913*
Energía eléctrica de Villa Clara	((5) 1 1)(1 1 0)4	Reg1991=*-14823.22174*
Petróleo de Villa Clara	(3 1 0)(2 1 0)4	Reg1991= *-3370.37009*

Se refleja de esta forma que también estadísticamente puede afirmarse que la provincia de Villa Clara fue afectada significativamente en los rubros de energía eléctrica y petróleo. Lo mismo se puede afirmar respecto al municipio de mayor cantidad de industrias en la provincia; concretamente Santa Clara.

4) Los modelos hallados para las series provinciales de consumo de petróleo y energía eléctrica, así como las series de consumo eléctrico de mayor envergadura en el municipio de Santa Clara, presentan similar comportamiento en las componentes regulares “p” y “q” respectivamente lo cual refleja la relación intima de estos rubros energéticos.
5) La presencia de regresores sólo en algunas series, así como la existencia de modelos diferentes, muestran que el impacto del período especial no se comportó de manera similar en estos municipios de la provincia de Villa Clara.
6) El uso de la constante en los modelos fue siempre considerada y sólo incluida cuando presentó significación estadística.
7) El uso del regresor “recuperación del período especial en el 2000” fue también considerado en algunas series que gráficamente pudieran sugerir alguna recuperación pero en ningún caso llegó a ser significativa.
8) Los pronósticos estimados fueron contrastados con los reales proporcionados por la Oficina Nacional de Estadística, y todos quedaron dentro de los intervalos de confianza y próximos la mayoría de ellos a los valores reales, demostrando la eficiencia de los modelos hallados.
9) A pesar de que los rubros consumo de energía eléctrica y consumo de petróleo son renglones influidos en sus mediciones por factores subjetivos, el trabajo realizado con las series mostró que las mismas son modelables de una forma suficientemente eficiente para que los pronósticos hallados fueran acertados, reflejando con ello su carácter objetivo, ratificado además por el claro comportamiento estacional de las mismas.

Para concluir el epígrafe, sólo resta destacar que en **el anexo III-14**, aparecen los pronósticos para el año 2006 y 2007 respectivamente, en el **anexo III-15,** un resumen de los modelos hallados, y en el **anexo III-16** los valores reales de consumo tanto de energía eléctrica como de petróleo reportados por la oficina Nacional de Estadística de Villa Clara para el año 2006.

3.5. Comparación de series de consumo de energía eléctrica en los municipios Caibarién y Placetas.

Como se observa anteriormente, varios municipios coincidieron en el modelo ARIMA hallado. En particular en Caibarién y en Placetas la introducción de un regresor para

modelar la influencia del período especial no fue significativa. Se decidió entonces comparar sus modelos siguiendo la metodología propuesta en el capítulo 1.

Las series de consumo eléctrico de ambos municipios partieron de 88 observaciones y condujeron ambas a modelos ARIMA (0,1,1)(0,1,1,)4. Por tanto, aparte de la estacionalidad de 4 (los datos son trimestrales), y las diferenciaciones regular y estacional, ambos modelos tienen dos parámetros MA1 (media móvil regular de orden 1) y SMA1 (media móvil estacional de orden 1). Comparemos estos modelos con la metodología descrita en el epígrafe 1.11. De los resultados del procedimiento de búsqueda del modelo tenemos los siguientes datos.

Estadísticas descriptivas de los coeficientes:

Para la serie de Caibarién	**N**	**Mean**	**Std**
MA1	88	0.7286	0.0769
SMA1	88	0.9249	0.1011
Para la serie de Placetas	**N**	**Mean**	**Std**
MA1	88	0.8485	0.0666
SMA1	88	0.3890	0.1110

Correlaciones entre los coeficientes:

Para la serie de Caibarién	**MA1**	**SMA1**
MA1	1	
SMA1	-0.0925	1
Para la serie de Placetas	**MA1**	**SMA1**
MA1	1	
SMA1	-0.3665	1

Calculamos entonces los valores conjuntos:

$$S_{MA1_p} = \sqrt{\frac{(0.0769)^2 + (0.0666)^2}{2}} \cong 0.0719$$

$$S_{SMA1_p} = \sqrt{\frac{(0.1011)^2 + (0.110)^2}{2}} \cong 0.1062$$

$$Corr_{MA1-SMA1_p} = \frac{(-0.0925) * 0.0769 * 0.1011 + (-0.3665) * 0.0666 * 0.1110}{2 * 0.0719 * 0.1062} \cong -0.2245$$

Entonces ya se puede preparar el fichero matricial con los siguientes datos:

ROWTYPE_	Municipio	VARNAME_	MA1	SMA1
N			176.0000	176.0000
MEAN	1		.7286	.9249
N	1		88.0000	88.0000
MEAN	2		.8485	.3890
N	2		88.0000	88.0000
STDDEV			.0719	.1062
CORR		MA1	1.0000	-.2245
CORR		SMA1	-.2245	1.0000

Aquí la variable municipio tiene las etiquetas 1: Caibarién 2: Placetas. Podemos ejecutar el commando MANOVA.

```
MANOVA ma1 sma1 BY Municipio(1,2)
/ PRINT=CELLINFO(MEANS) ERROR
/ MATRIX=IN(*).
```

Los resultados esenciales son los siguientes (*comentarios en italics*).

```
* * * * * * A n a l y s i s   o f   V a r i a n c e * * * * * *
       176 cases accepted.
         0 cases rejected because of out-of-range factor values.
         0 cases rejected because of missing data.
```

Las estadísticas descriptivas permiten comprobar que los datos han sido bien captados

```
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Cell Means and Standard Deviations
 Variable .. MA1
      FACTOR          CODE                Mean  Std. Dev.          N
  Municipi        Caibarié                .729       .            88
  Municipi        Placetas                .849       .            88
 For entire sample                        .789       .           176
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Variable .. SMA1
      FACTOR          CODE                Mean  Std. Dev.          N
  Municipi        Caibarié                .925       .            88
  Municipi        Placetas                .389       .            88
 For entire sample                        .657       .           176
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 WITHIN CELLS Correlations with Std. Devs. on Diagonal
                  MA1       SMA1
 MA1             .072
 SMA1           -.225       .106
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

El test de esfericidad descarta que la matriz de correlaciones sea la identidad y por tanto el enfoque multivariado es apropiado.

```
 Statistics for WITHIN CELLS correlations
 Log(Determinant) =                -.05171
 Bartlett test of sphericity =    8.92079 with 1 D. F.
 Significance =                      .003
 F(max) criterion =               2.18168 with (2,174) D. F.
```

Los resultados del Análisis de Varianza multivariado demuestran que en su conjunto los coeficientes de las dos series difieren.

```
* * * * * * A n a l y s i s   o f   V a r i a n c e * * * * * *
 WITHIN CELLS Sum-of-Squares and Cross-Products
                  MA1       SMA1
 MA1             .900
 SMA1           -.298      1.962
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

* * * * * * A n a l y s i s   o f   V a r i a n c e * * * * * *
 EFFECT .. Municipio
 Multivariate Tests of Significance (S = 1, M = 0, N = 85 1/2)

 Test Name       Value    Exact F Hypoth. DF   Error DF  Sig. of F
 Pillais          .86694  563.56450       2.00     173.00       .000
```

```
Hotellings      6.51520  563.56450       2.00     173.00        .000
Wilks            .13306  563.56450       2.00     173.00        .000
Roys             .86694
Note.. F statistics are exact.
```

Los resultados del Análisis de Varianza de cada coeficiente por separado, demuestra también que ambos difieren entre los municipios

```
EFFECT .. Municipio (Cont.)
Univariate F-tests with (1,174) D. F.

Variable   Hypoth.SS   Error SS Hypoth.MS   Error MS      F  Sig.of F
MA1          .63254     .89951     .63254     .00517  122.35825   .000
SMA1       12.63631    1.96245   12.63631     .01128 1120.39499   .000
```

Se debe descartar la igualdad de los modelos de consumo energético para Caibarién y Placetas. Para interpretar esta conclusión se puede acudir a los estadígrafos generales de ambos procesamientos. Si bien en ambos, los dos coeficientes: MA1 y SMA1 son significativos, el role de los mismos en cada municipio es ligeramente diferente. En el caso de Caibarién, ambos coeficientes tienen magnitudes relativamente similares (0.7286 y 0.9249) y también similares significaciones. De hecho véase que las razones del test de Student en Caibarién para estos coeficientes son de 9.4789 y 9.1464. En cambio en el modelo de Placetas, es mucho más marcado el role de MA1: 0.8485 (con T-Ratio 12.7363) que el de SMA1: 0.3890 (T-Ratio 3.5038) y significaciones diferentes. Es más, en el municipio de Placetas la correlación entre MA1 y SMA1 es mucho más marcada (-0.3665) que en Caibarién (-0.0925). En otras, palabras, el consumo en Caibarién depende claramente de los cambios en el consumo del trimestre anterior y los cambios en el consumo del año anterior. En el caso de Placetas, la dependencia está más centrada en el trimestre anterior, la incidencia del cambio en el año anterior está menos marcada y en cierta forma abarcada por el cambio de los trimestres. Definitivamente Placetas “actualiza” su consumo mucho más dinámicamente que Caibarién.

Las ideas aquí expuestas de trabajo con matrices pueden ser aplicadas en situaciones mucho más generales. Ver si se quiere la ayuda del SPSS sobre el comando MANOVA y en particular sobre el subcomando MATRIX dentro de MANOVA Univariate (lo cual se hace extensivo a MANOVA Multivariate y otras formas de MANOVA**), ver Anexo III-17**.

Conclusiones parciales

En este capítulo se muestran los resultados principales del trabajo. Se obtuvieron modelos ARIMA de consumo energético para la provincia y para los cuatro municipios estudiados. Se obtuvo también el modelo para la serie de consumo de petróleo provincial. En todos los casos se intentó modelar la influencia negativa del período especial y su posible recuperación. En las series provinciales y en el municipio más industrializado (Santa Clara) la introducción de un regresor en el año 1990 fue significativo, no sucedió así en los otros tres municipios analizados.

El intento de modelación de una recuperación brusca fue fallido, pues en ninguno de los casos la introducción del regresor fue significativa. Sin embargo se hizo un estudio de la serie de consumo de petróleo provincial en el período 1993–1999, se probó que en este período su tendencia es creciente, pero ello no tiene una incidencia significativa sobre el comportamiento posterior de la serie, por lo cual se puede hablar apenas de una recuperación paulatina.

Por último se aplicó la metodología descrita en el capítulo 1 para comparar modelos de municipios diferentes. Se utilizaron los resultados obtenidos para Placetas y Caibarién. Se llegó a la conclusión de que, a pesar de que ambos coinciden (ARIMA(0, 1, 1)(0, 1, 1)4) sus coeficientes difieren significativamente, por lo que el comportamiento del gasto energético en ambos municipios es diferente.

Conclusiones

El estudio detallado de los datos periódicos sobre el consumo de petróleo y de energía eléctrica, proporcionados por la Oficina Nacional de Estadística de Villa Clara, ha posibilitado la modelación matemática de estos fenómenos. Basado en los resultados obtenidos, se enuncian las siguientes conclusiones:

1. Se obtuvieron modelos ARIMA de las series de consumo eléctrico de los municipios Caibarién, Placetas, Sagua y Santa Clara.
2. Se modeló matemáticamente la serie de consumo de petróleo y de energía eléctrica en la provincia de Villa Clara. En ambas se introdujo un regresor (que resultó ser significativo) para modelar la influencia del período especial. Se introdujo otro regresor de pulso instantáneo (que resultó ser no significativo) para caracterizar el proceso de recuperación.
3. La serie de energía eléctrica del municipio de Santa Clara se modeló desde el punto de vista clásico y siguiendo la metodología de Box – Jenkins. Se compararon ambos modelos y se mostró la superioridad del último.
4. Se aplicó con éxito la metodología general para comparar modelos ARIMA. En particular se compararon los modelos de los municipios de Caibarién y Placetas y se encontraron diferencias entre ambos.
5. Los pronósticos realizados para el año 2007 a partir de los modelos obtenidos se ajustan aceptablemente según criterios de la Oficina Nacional de Estadística

Recomendaciones

- Realizar el análisis de otras series de rubros económicos para su modelación y pronóstico.
- Desarrollar métodos estadísticos de comparación de modelos para una misma serie en el tiempo, o sea técnicas que identifiquen el momento en el que una serie dada cambia su modelo inicial.

Bibliografía

Akaike, H. (1974). "A New look at Statistical Model Identification." IEEE Transaction on Automatic Control **Ac-19**: 718-723.

Arellano, M. (2006). "Introducción al análsis Clásico de series de Tiempo." from http://www.ciberconta.unizar.es/LECCION/SERIES.

Box, G. a. T., G. (1975). "Intervention analysis with application to economic environmental problems." Journal of the American Statistical Association. **70**: 70-79.

Box, G. E. P. a. J., G.M. (1994). Time Series Analysis Forecasting And Control. . San Francisco, Holden- Day.

Cochrane, J. H. (1997). Time Series for Macroeconomics and Finance Chicago, University of chicago.

Cué Muñiz, J. E. C. E. (1987). Estadística.

Diebold , F. X. (2000). Elements of Forecasting. Pennsylvania, University of Pennsylvania.

Fuller, W. (1976). Introduction to Statistical Time Series New York, Wiley Series in Probability and Mathematical Statistic. John Wiley and Sons

Gladys Casas, R. G., Milagros Alegret (1999). Métodos para la vigilancia de eventos (III): Técnicas de Clustering para la Detección de Epidemias.Reporte Técnico de Vigilancia ,julio,1999,4(7). Ciencias de la Computación, uclv. **Master en Ciencias**.

Granma, p. (2007). cuba.

Grau, A. R. (1994). Estadística Aplicada con ayuda de paquetes de software. Universidad guadalajara, Jalisco, Mexico.

Grau, A. R. (1996). Series Cronológicas, Curso de Especialización en Procesos Estadísticos Aplicados. Colombia, Coruniversitaria, Ibagué.

Guerrero, V. M. (1991). Análisis Estadístico de series de tiempo Económicas. México, Colección CBI.Universidad Autónoma Metropolitana.
.

Gupta, V. (1999). SPSS for Beginners, VJBooks Inc.

Jeffrey, W. H. a. B., J. O (1992). "Ockham's Razor and Bayesian Analysis. ." Am. Sci **80**: 64-72.

Jeffrey, W. H. a. B., J.O (1992). "Ockham's Razor and Bayesian Analysis." Am. Sci **80**: 64-72.

Koroliov, V. (1986). Manual de la teoría de probabilidades y estadística matemática.

Medina, J. H. (1998). Estudio del comportamiento histórico de las tasas de las enfermedades de declaracion obligatoria(EDO) en el municipio de Manicaragua. Santa Clara Villa Clara, Universidad Central De Las VIllas. **Maestria**.

Mondeja Hernandez, A. L. (1995). Metodología para el uso de las series de tiempo en epidemilogía. Santa Clara, Villa Clara, UCLV. **Master en Ciencias:** 98.

Monteagudo, M. P. (2007). Series de Cronológicas de lluvia en la cuenca sagua la chica. Modelos y pronósticos. matematica. santa clara, Villa Clara, Universidad Central de Las Villas. **Diploma**.

Mora Villegas, H. (2003). Series crónologicas de consumo eléctrico y de petróleo de los municipios y provincia de Villa Clara Santa Clara, Villa Clara., UCLV. **Master en Ciencias.:** 105.

Osés Rodríguez, R. (2004). Series Meteorológicas de Villaclara y otras provincias. Modelos y Pronósticos. Santa Clara,Villa Clara, UCLV. **Master en Ciencias:** 98.

Peña, D. S. d. R. (1999). Estadística, Modelos y Métodos. Madrid.

Schwartz, G. (1976). "Estimating the dimensions of a model. ." Annals of Statistic **6**: 461-464.

Shumway, R., Stoffer D. (2000). Time Series Analysis and its Applications. Pittsburgh, University of Pittsburgh.

Tarrau Brito, M. E. (1996). Caracterización de las series cronológicas de enfermedades diarreicas y respiratorias agudas en Villa Clara. Santa Clara, Villa Clara., UCLV. **Master en Ciencias:** 115.

Tiao, C. G. a. T., R.S, (2001). A Course in Time Series Analysis. New York., John Wiley

Anexos