

1 03043



UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

INSTITUTO DE INVESTIGACIONES EN MATEMATICAS APLICADAS Y EN SISTEMAS

“UN MODELO DE SERIES DE TIEMPO PARA EL PORCENTAJE DE OCUPACION EN CANCUN”

T E S I S I N A
QUE PARA OBTENER EL DIPLOMA DE
ESPECIALIZACION EN ESTADISTICA
A P L I C A D A
P R E S E N T A :
RICARDO HERNANDEZ BARAJAS

DIRECTOR: M. EN C. JOSE SALVADOR ZAMORA MUÑOZ



2002

TESIS CON FALLA DE ORIGEN



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



Como siempre,

*gracias a **DIOS***

y

gracias a mis padres Irma y Francisco.





A mi flaquita (Maribell), gracias por ser protagonista de mis silencios, de mis risas, de mi realidad y de mis sueños. A ti

*Mi Yo poeta te diría que
"soñar es sujetar el agua,
esculpir el humo,
tatuarse el aire,
dibujar un pensamiento".*

*Mi Yo matemático agregaría:
"es cruzar líneas paralelas
construir conjuntos sin medida,
encontrar infinitos cada vez mayores,
y resolver una integral sin límite".*

*Finalmente, mi Yo enamorado sólo te dice:
"es tener la certeza de que existes,
y al mismo tiempo, estás a mi lado;
es imaginarte por siempre
con mi rostro reflejado en tus ojos color miel".*

Te amo trompuda. Gracias por ser y por estar al mismo tiempo.





Gracias...



*Por su paciencia y apoyo,
a mis hermanos Paco, Marco y Clau.*



*Porque tengo lo más valioso de ellos, su amistad,
a Ita, Luisa, Norma y Mario.*

También por su amistad a:

*Arturo, Blanquita, Brisa, Don Fer, Fabiola, Felix,
Moy, Paty, Perla, Raquel, Sandy, Tony..., y los que me
faltaron, pero que estoy seguro sentirán gusto al saber
que he conseguido una meta más en mi vida.*



*Por sus consejos, su tiempo, su atención, su paciencia, su apoyo y su buen
humor, pero sobre todo por su forma de dar clase con todas esas cualidades
a Salvador.*

*También por su tiempo y apoyo al revisar este trabajo, a mis sinodales:
Leticia Gracia, Rebeca Aguirre, Alberto Contreras y Carlos
Díaz.*



ÍNDICE

Contenido	Pág.
RESUMEN	1
INTRODUCCIÓN	2
EL PORCENTAJE DE OCUPACIÓN COMO INDICADOR	2
PLANTEAMIENTO DEL PROBLEMA	3
SERIES DE TIEMPO COMO ALTERNATIVA	3
Objetivo	3
ZONA DE ESTUDIO	4
CAPÍTULO 1. CONCEPTOS DE SERIES DE TIEMPO	5
1.1. PROCESOS ESTOCÁSTICOS Y SERIES DE TIEMPO	5
1.2. NOTAS SOBRE EL USO DE SERIES DE TIEMPO	6
1.3. MOMENTOS	6
1.4. FUNCIONES DE AUTOCORRELACIÓN FAC Y FACP	7
1.5. DESCOMPOSICIÓN DE LA VARIABILIDAD	7
1.5.1. Teorema de Descomposición de Wold	8
1.6. EL CORRELOGRAMA	8
1.6.1. Cálculo de la FAC y la FACP Muestrales	10
CAPÍTULO 2. MODELOS	11
2.1. RUIDO BLANCO	11
2.2. PROMEDIOS MÓVILES (MA)	12
2.2.1. Notas sobre los Modelos MA	12
2.3. AUTOREGRESIVO (AR)	13
2.3.1. Notas sobre los Modelos AR	13
2.4. ARMA	14
2.4.1. Notas sobre los Modelos ARMA	14
2.5. ARIMA	14
2.5.1. Operadores Diferencia (∇) y Retraso (B) para la Transformación de Series No-Estacionarias	15
2.5.2. Forma de un Modelo ARIMA	15
2.5.3. Notas sobre los Modelos ARIMA	16
2.5.4. ARIMA para Series Periódicas	16
2.5.5. Predicción con ARIMA	16
CAPÍTULO 3. SELECCIÓN DEL MODELO	18
3.1. OBTENCIÓN DE UNA SERIE ESTACIONARIA	18
3.1.1. Estabilización de la Varianza	19

ÍNDICE

(Continuación)

Contenido	Pág.
3.1.2. Estabilización de la Media	21
3.2. ETAPA 1: IDENTIFICACIÓN	21
3.3. ETAPA 2: ESTIMACIÓN	23
3.4. ETAPA 3: VERIFICACIÓN	23
3.4.1. Supuesto 1: Toda Z_t posee Distribución Normal	24
3.4.2. Supuesto 2: Toda Z_t posee Media Cero y Varianza Constante	24
3.4.3. Supuesto 3: Las Z_t 's son No-correlacionadas	25
3.5. OTRAS CONSIDERACIONES	25
3.5.1. Análisis en el Dominio de las Frecuencias	26
CAPÍTULO 4. ANÁLISIS DE LA OCUPACIÓN HOTELERA EN CANCÚN	28
4.1. TRANSFORMACIÓN A UNA SERIE ESTACIONARIA	28
4.1.1. Estabilización de la Varianza	29
4.1.2. Eliminación del Comportamiento Periódico	31
4.1.3. Eliminación de Tendencia	33
4.2. IDENTIFICACIÓN	34
4.3. ESTIMACIÓN	36
4.4. VERIFICACIÓN	37
4.4.1. Normalidad de Residuos	39
4.4.2. Residuos con Media Cero y Varianza Constante	39
4.4.3. Residuos No-correlacionados	40
4.4.4. Análisis en el Dominio de las Frecuencias	41
4.5. PREDICCIONES	43
CAPÍTULO 5. DISCUSIÓN FINAL	45
5.1. RESULTADOS	45
5.2. CONCLUSIONES	46
ANEXOS	49
A.1. CUESTIONARIO HOTELERO	49
A.2. CÁLCULO DE LA FAC Y FACP CON PAQUETERÍA	51
A.3. AJUSTE DE UN MODELO ARIMA EN SPSS	54
A.4. MÉTODO BOX-COX EN S-PLUS	59
A.5. CÁLCULO DE PERIODOGRAMAS EN SPSS	60
BIBLIOGRAFÍA	62

RESUMEN

¿La ocupación hotelera es un fenómeno puramente aleatorio, completamente determinístico, o una combinación de ambos? En caso de no ser sólo azar, ¿puede la ocupación ser pronosticada a partir de sus valores pasados?; y de ser así, ¿depende de la información inmediata, de la del año anterior, o de toda la anterior?

El presente trabajo, utiliza un modelo de series de tiempo para responder a estas y algunas otras preguntas de interés, sobre todo, para los sectores involucrados con el turismo. Su principal contribución, además de modelar y pronosticar los porcentajes de ocupación, es proporcionar intervalos de confianza para las proyecciones obtenidas.

Algunas conclusiones obtenidas para el fenómeno de porcentajes de ocupación mensuales en Cancún, fueron que:

- i) Existe un comportamiento periódico, ya que, en cada lapso de doce meses los datos se presentan de forma similar al periodo inmediato anterior.
- ii) La ocupación puede explicarse con base en la información histórica (pasada), utilizando un modelo $ARIMA(3,1,0)_{12}$ para los inversos de los porcentajes.

El estudio se divide en tres partes, la primera, formada por el capítulo introductorio, muestra los aspectos generales que originaron el planteamiento del problema, los objetivos perseguidos para su solución, y la selección de una alternativa que permitiera que éstos fueran alcanzados.

La segunda parte, constituida por los capítulos 1 a 3, presenta los conceptos, modelos, técnicas y herramientas necesarias para el ajuste de un modelo de series de tiempo.

En la última parte (capítulos 4, 5 y anexos), se modela el porcentaje de ocupación hotelera mensual en Cancún, y se realiza un esfuerzo por conseguir que, usando paquetería estadística, el lector pueda reproducir lo obtenido; además de que logre adecuarlo o modificarlo para el análisis de otras variables, en particular, las manejadas por el sector, tales como la llegada de visitantes y los ingresos por actividades turísticas.

INTRODUCCIÓN

El crecimiento sostenido del movimiento de visitantes con fines turísticos, entre los países y al interior de los mismos, es una característica del fin del siglo pasado y principios del actual. Este fenómeno exige una mayor oferta de servicios turísticos, al mismo tiempo que potencia al sector como un importante generador de divisas y empleo.

“En la mayoría de los países, las actividades de servicios tienen una importancia creciente como fuente de ingresos y empleo. Dentro de los servicios, los relacionados con el turismo representan un papel importante, en particular, en los países o regiones que tienen algún atractivo para los usuarios, debido por ejemplo, al clima, a las bellezas naturales, a los lugares históricos o a precios ventajosos. Para algunos de ellos, el turismo es una fuente de recursos exteriores que tiene una importancia crucial en su desarrollo económico”.

“No obstante, la capacidad para atraer un flujo permanente de turismo exige el diseño de un conjunto completo de medidas, que abarca desde la construcción de infraestructuras de transporte y alojamiento que alcancen ciertos niveles aceptables internacionales, hasta la conservación y preservación de los lugares que constituyen la principal atracción turística, una política de promoción y comercialización, la formación de una mano de obra adecuada, la educación de la población para conservar el medio ambiente, etc. ...”, (ALADI [3]).

EL PORCENTAJE DE OCUPACIÓN COMO INDICADOR

Para cuantificar los beneficios de la actividad turística es imprescindible obtener, procesar y predecir información con la periodicidad requerida por los usuarios; los cuales tienen necesidades como el análisis de mercado, la evaluación de la eficacia de las campañas de comercialización, las inversiones en la industria, el desarrollo de los recursos humanos, la toma de decisiones en materia de política turística y de ecología, o simplemente la defensa de intereses particulares.

Uno de los datos estadísticos que la Organización Mundial de Turismo (OMT) recomienda sea monitoreado es el **porcentaje o tasa de ocupación** en los establecimientos de hospedaje, y en particular, en los hoteles. Dado que el presente trabajo estudia este tipo de información, en la figura I.1 se incluye la forma en que se calcula una tasa de ocupación mensual.

En el caso de México, estos porcentajes son obtenidos por cada una de las Oficinas Estatales de Turismo en coordinación con la Secretaría de Turismo (SECTUR), a través del instrumento conocido como Sistema de Información Turística Estatal (SITE), donde se incluye una opción llamada “**Programa Cuestionario Hotelero**”. En este cuestionario, (véase el anexo A.1) se captura la información diaria de una muestra de hoteles (datos como

la llegada de huéspedes, los cuartos ofertados y los ocupados). Finalmente, con la información de la muestra, y usando la fórmula ya descrita, el SITE calcula los porcentajes de ocupación para cada localidad turística.

$$\% \text{ de Ocupación del Mes} = \frac{(\text{Cuartos Ofertados Diariamente}) * (\text{Días del Mes})}{\sum_{i \in \{\text{Días del Mes}\}} (\text{Cuartos Ocupados del Día } i)} * 100$$

Figura I.1. Fórmula para el cálculo del porcentaje de ocupación mensual de una localidad.

PLANTEAMIENTO DEL PROBLEMA

Una gran limitante de los datos estadísticos procesados por la SECTUR, incluidos los porcentajes de ocupación, es que sólo son generados por medición de lo observado. *¿Cómo hacer una real planeación de las políticas turísticas si no se cuenta con los posibles escenarios que presentarán los datos en el futuro?*

El objetivo general de este trabajo es ejemplificar que con herramientas estadísticas, se pueden modelar fenómenos como la ocupación, y al mismo tiempo, a través de esos modelos, hacer predicciones sobre su comportamiento futuro. Más específicamente, *se pretende plantear de forma sencilla como “modelar” el porcentaje de ocupación mensual de Cancún, Quintana Roo; poder predecir porcentajes futuros y, saber “qué tan buenas” son las predicciones.* Todo lo anterior, sin la restricción de tener grandes conocimientos en Matemáticas o Estadística. Por último, debe aclararse que con “modelar”, también se hace referencia al hecho de saber cuándo el modelo ajustado simula aceptablemente la realidad del fenómeno.

SERIES DE TIEMPO COMO ALTERNATIVA

De manera general, una **serie de tiempo** es un conjunto de observaciones-mediciones de un fenómeno, hechas secuencialmente en el tiempo. Existe una gran variedad de ejemplos de series de tiempo (y no sólo en la literatura), los cuales van desde los económicos y los demográficos, hasta los físicos y los biológicos. Ahora, dado que la ocupación hotelera en Cancún es un fenómeno observado y medido mensualmente, entonces puede ser visto como una serie de tiempo; por lo cual, *se tomó la decisión de modelarle con esta técnica.*

Objetivo

Una vez que se decidió utilizar las técnicas de Series de Tiempo para solucionar el problema planteado, fue posible enunciar el objetivo principal de este trabajo, de forma que,

al alcanzarse, cubriera también los objetivos que de forma general, se habían mencionado en la sección anterior:

Ajustar un modelo de series de tiempo a los porcentajes de ocupación mensual de Cancún, Quintana Roo, y, evaluar qué tan apropiado es el ajuste para la explicación y para la predicción del fenómeno.

ZONA DE ESTUDIO

En 1967, el gobierno de México realizó un estudio con el fin de localizar el mejor lugar para construir un complejo turístico moderno, lujoso y redituable que aprovechara la belleza del paisaje caribeño del país. En 1970 se tomó la decisión: **Cancún, situado en el estado de Quintana Roo a 1,722 kilómetros al este de la Ciudad de México** (figura I.2).

Aunque en sus inicios el lugar estaba prácticamente deshabitado, sin caminos ni los más elementales servicios como agua potable, actualmente posee a lo largo de más de 20 kms. de litoral, alrededor de 400 restaurantes, grandes centros comerciales y más de 400,000 habitantes. Con respecto al hospedaje, existen 127 establecimientos, de los cuales **102 son hoteles con capacidad total de 21,664 cuartos** ofertados diariamente a los visitantes. Es precisamente sobre estos cuartos ofertados que el presente trabajo analiza el comportamiento del porcentaje de ocupación mensual.

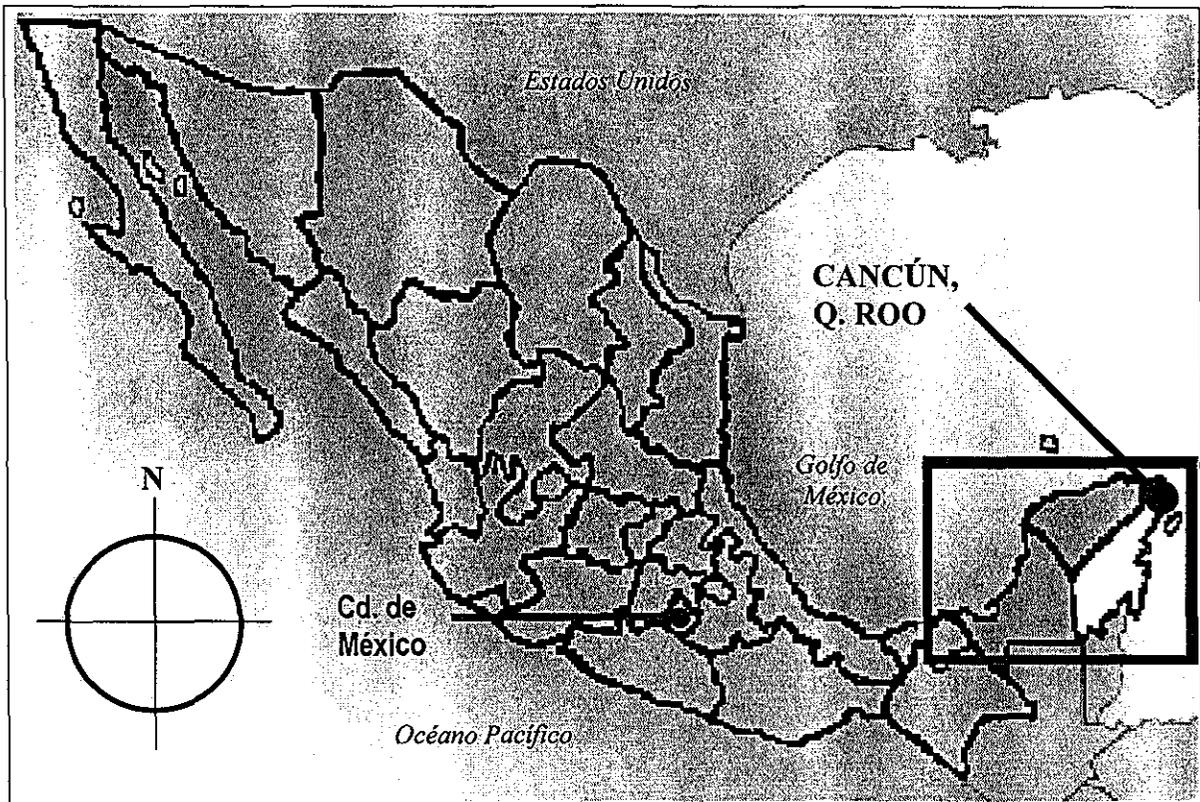


Figura I.2. Ubicación del área de estudio.

Capítulo 1

CONCEPTOS DE SERIES DE TIEMPO

A lo largo de este capítulo se introducen las definiciones y los conceptos principales de la teoría de Series de Tiempo, herramientas básicas para comprender la forma en que se ataca el problema de modelar el fenómeno de porcentajes de ocupación. Para mayor detalle sobre los conceptos y ejemplos, puede consultarse Chatfield [3] o Box & Jenkins [6].

1.1. PROCESOS ESTOCÁSTICOS Y SERIES DE TIEMPO

Chiang [7] define un **proceso estocástico** como un fenómeno que se desarrolla a través del tiempo, siguiendo leyes probabilísticas. Así, dada la idea intuitiva descrita en el capítulo introductorio, es fácil observar que una **serie de tiempo** es un caso particular de un proceso estocástico (de hecho, es la realización de uno).

Las series de tiempo se dividen en dos grandes grupos: las **continuas** y las **discretas**. Se dice que una serie es continua cuando las observaciones se hacen permanentemente en el tiempo; mientras que se le llama discreta a aquella en que las observaciones se toman en momentos específicos y que generalmente son igualmente espaciados.

Dada la división anterior y las características del fenómeno estudiado (porcentajes de ocupación mensuales), es claro que en el presente trabajo, se emplearán las técnicas existentes para el *análisis de series de tiempo discretas*, en donde las *observaciones son igualmente espaciadas* en el tiempo.

1.2. NOTAS SOBRE EL USO DE SERIES DE TIEMPO

Contrariamente a la gran mayoría de métodos y herramientas estadísticas utilizadas en la práctica, el análisis de series de tiempo *no supone independencia entre observaciones*; y no sólo eso, sino que además *usa la dependencia para la modelación del fenómeno*, lo cual constituye una ventaja más para el problema de explicación y predicción de los porcentajes de ocupación

Con respecto a esto último, es de vital importancia aclarar que, *la mayoría de las series son estocásticas*; y por lo tanto, su futuro sólo se determina parcialmente por su pasado. Sin embargo, también debe recordarse que en la estadística, las predicciones tienen asociada una distribución de probabilidad, la cual permite *construir intervalos de confianza para estas predicciones*.

1.3. MOMENTOS

Las principales metas al realizar un análisis de series de tiempo son: la *descripción* del fenómeno, su *explicación*, la *predicción* de valores futuros, y el *control* de su comportamiento (como el control de calidad en la producción de un objeto).

Si se denota a una serie de tiempo como $\{X(t)\} = \{X_t\}$ (donde t es el tiempo y X_t es la observación), una forma de conseguir la descripción de cualquier proceso estocástico, y por lo tanto de una serie de tiempo, es a través de su función de distribución de probabilidad conjunta $F(X(t_1), \dots, X(t_n))$; sin embargo, es muy complicada. Otra forma de describirles, que es la más usada, es a partir de sus **momentos**, los cuales se definen en la tabla 1.1.

Media	$\mu(t) = E[X(t)]$
Varianza	$\sigma^2(t) = \text{Var}[X(t)]$
Autocovarianza	$\gamma(t_1, t_2) = E\{ [X(t_1) - \mu(t_1)] * [X(t_2) - \mu(t_2)] \}$

Tabla 1.1. Momentos de una serie de tiempo.

Cuando el **proceso es estacionario** (esto es, cuando no hay cambios sistemáticos ni en la media ni en la varianza a través del tiempo, y cuando no tiene comportamientos periódicos), la autocovarianza puede reescribirse como

$$\gamma(\tau) = \text{Cov} [X(t), X(t + \tau)] ,$$

esto es, que *la correlación no depende del tiempo*.

Ya que se mencionó a los procesos estacionarios, es posible definir además, el término de **serie estrictamente estacionaria**, que se refiere a aquellas series que cumplen:

$$\forall n, \tau: F(X(t_1), \dots, X(t_n)) = F(X(t_1 + \tau), \dots, X(t_n + \tau)) .$$

En otras palabras, *una serie es estrictamente estacionaria si su comportamiento es el mismo al trasladarse en cualquier magnitud de tiempo* (τ).

1.4. FUNCIONES DE AUTOCORRELACIÓN FAC Y FACP

Dado que la magnitud del coeficiente de autocovarianza depende de las unidades en que se mide X_t , usualmente se utiliza la **función de autocorrelación (FAC)**, que se obtiene de estandarizar la autocovarianza, y que se calcula como:

$$\rho(t) = \gamma(t) / \gamma(0)$$

La FAC se interpreta como la *correlación entre observaciones de la misma variable a diferentes tiempos* (entre X_t y $X_{t+\tau}$).

Otra función que mide la correlación entre observaciones X_t y X_{t+k} es la **función de autocorrelación parcial (FACP)**, que *cuantifica la correlación tomando en cuenta sus observaciones intermedias* $X_{t+1}, \dots, X_{(t+k)-1}$. Lo anterior se logra con la siguiente expresión:

$$\varphi_k = \frac{\text{Cov}[(X_t - \hat{X}_t), (X_{t+k} - \hat{X}_{t+k})]}{\sqrt{\text{Var}(X_t - \hat{X}_t)} \sqrt{\text{Var}(X_{t+k} - \hat{X}_{t+k})}},$$

donde

$$\hat{X}_{t+k} = E\{X_{t+k} | X_{t+1}, \dots, X_{(t+k)-1}\}$$

es la esperanza condicional y estima la dependencia lineal de X_{t+k} con $X_{t+1}, \dots, X_{(t+k)-1}$.

Es de resaltar que *estas funciones son generalmente estimadas con paquetería* (véase el anexo A.2); por lo cual, aún las personas poco familiarizadas con la teoría estadística pueden aprovechar las técnicas de series de tiempo. Basta con entender qué representa cada coeficiente.

1.5. DESCOMPOSICIÓN DE LA VARIABILIDAD

Al realizar un análisis de series de tiempo, y luego de investigar las características de los datos, y tener bien definidos los objetivos de un estudio, es importante graficar las observaciones contra el tiempo ("t vs X_t "). Con esto, se pueden vislumbrar, entre otras cosas, las posibles tendencias, estacionalidades o ciclos, discontinuidades y datos atípicos; por lo tanto, esta gráfica es fundamental en la búsqueda de un modelo que describa adecuadamente al fenómeno de estudio.

Tradicionalmente, al realizar un análisis de series de tiempo se intenta explicar la variabilidad del fenómeno con base en cuatro efectos:

1. *Efecto Estacional*. Es una variación periódica de amplitud constante; por ejemplo, semanal, anual o mensual como en el caso que se estudiará de porcentajes de ocupación.
2. *Cambio Cíclico*. Es la variación de la serie que tiene amplitud no constante; por ejemplo, en Economía, además de la variación anual, suelen presentarse ciclos de auge o depresión, que son predecibles pero que pueden no tener espacios de tiempo iguales entre ellos.
3. *Tendencia*. Es un cambio suave a través del tiempo en el valor medio de los datos; por ejemplo, en la llegada mensual de turistas extranjeros a México, aparte de presentarse oscilaciones a lo largo de un año o ciclos de auge a lo largo del tiempo, ha ido creciendo, entre otras razones, por el aumento de los turistas, el de las opciones turísticas del país, y conforme se ha desarrollado la tecnología en los medios de transporte.
4. *Fluctuación Irregular o Aleatoria*. Es el tipo de variación que queda en una serie, al sustraerle tanto los efectos estacionarios, como los cambios cíclicos y las tendencias; esto es, la variación causada por situaciones no controlables o predecibles (lo aleatorio de la serie).

1.5.1. Teorema de Descomposición de Wold

Un resultado de gran interés, sobre todo para la comprensión de la serie estudiada en este trabajo, es el teorema de descomposición de Wold [9], el cual afirma que:

Cualquier proceso estacionario discreto, puede ser expresado como la suma de dos procesos no correlacionados, uno completamente determinístico y otro puramente aleatorio.

1.6. EL CORRELOGRAMA

El **correlograma** es la gráfica que contiene a *los valores estimados para la FAC (valores muestrales)*, denotados como **coeficientes de autocorrelación** r_k ; esto es, “k vs. r_k ”. La inspección visual de éste, resulta de gran ayuda al proponer el modelo que debe ser ajustado, ya que puede detectarse si la serie tiene tendencia, efectos estacionales o ciclos; o bien, si carece de ellos y es completamente aleatoria. Además, es posible observar si existe “correlación a corto plazo” (explicada más adelante).

La interpretación del correlograma no suele ser fácil, a continuación se describen algunos lineamientos generales proporcionados por Chatfield [1].

No-Correlación. Se presenta cuando los coeficientes de autocorrelación son cercanos a cero; esto es, cuando

$$\forall k \neq 0, r_k \cong 0 .$$

De hecho, si la serie es de observaciones no-correlacionadas, entonces los coeficientes de autocorrelación se distribuyen asintóticamente como una normal de media 0 y varianza $1/N$ (donde N es el número de observaciones); es decir:

$$r_k \overset{a}{\sim} N(0, 1/N) .$$

Correlación a Corto Plazo. Ocurre cuando los coeficientes de correlación tienden a ser cada vez más pequeños ($|r_k| \rightarrow 0$), y por lo tanto, la correlación entre los valores observados decrece conforme más separados estén en el tiempo.

Alternancia. Cuando la serie tiende a alternar (mostrando observaciones “arriba-abajo-arriba” de una media global), el correlograma suele presentar el mismo comportamiento alrededor del cero.

Tendencia. En el caso en que la serie posee una tendencia, los valores r_k tienden a cero mucho más lentamente que cuando sólo se tiene aleatoriedad.

Fluctuaciones Estacionales o periódicas. Si existe este comportamiento en la serie, el correlograma muestra oscilaciones periódicas; por ejemplo, puede suceder que con observaciones mensuales, r_6 (correlación estimada entre las observaciones X_t y X_{t+6}) sea “grande y negativo”, mientras que r_{12} “grande y positivo”. Es importante observar que al “remover” la variación estacional de una serie, el nuevo correlograma puede aún dar más información, como aleatoriedad o correlación a corto plazo.

Observaciones Atípicas. Cuando la serie presenta observaciones poco comunes o “raras”, el correlograma obtenido al ajustarles puede diferir mucho del original, por lo cual es conveniente detectarles antes. Una técnica para lograrlo es graficar “ x_t vs. x_{t+k} ” y observar si existen puntos extremos.

1.6.1. Cálculo de la FAC y la FACP Muestrales

En la práctica, los coeficientes de las funciones de autocorrelación y de autocorrelación parcial (FAC y FACP) se calculan utilizando los **coeficientes de autocovarianza muestral**, los cuales se definen, para un retraso k , como:

$$C_k = \sum_{t=1}^{N-k} \frac{(x_t - \bar{x})(x_{t+k} - \bar{x})}{N},$$

donde x_t es el valor observado y \bar{x} la media muestral.

A partir de estos coeficientes se obtiene un estimador para $\rho(k)$, que es

$$\hat{\rho}(k) = r_k = \frac{C_k}{C_0}$$

Ya calculados los r_k , los coeficientes de autocorrelación parcial se estiman de la siguiente forma:

$$\hat{\phi}_{1,1} = r_1$$

$$\hat{\phi}_{k,k} = \frac{r_k - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_j}, \quad \text{para } k > 1;$$

donde

$$\hat{\phi}_{k,j} = \hat{\phi}_{k-1,j} - \hat{\phi}_{k,k} \hat{\phi}_{k-1,k-j}, \quad \text{para } k > 2.$$

(Para el cálculo con paquetería de la FAC y la FACP, puede consultarse el anexo A.2; mientras que si se desea profundizar en la obtención de los estimadores, puede consultarse [1].)

Capítulo 2

MODELOS

En este capítulo se presentan cuatro tipos de modelos de series de tiempo, que van desde el más simple, conocido como “ruido blanco”, hasta el más complicado y que generaliza a los otros, el modelo ARIMA. Su descripción tiene como objetivo facilitar y comprender cuál es conveniente seleccionar para su ajuste durante el estudio de un fenómeno.

2.1. RUIDO BLANCO

Un **ruido blanco**, es una sucesión de variables aleatorias, $\{Z_t\}$, todas *no-correlacionadas* entre sí y con *igual distribución* de probabilidad.

Por ser idénticamente distribuidas, es claro que el proceso tiene *media y varianza constantes* y que, por ser independientes, si $k \neq 0$, entonces

$$\gamma(k) = \text{Cov}(Z_t, Z_{t+k}) = 0.$$

Muy importante es el hecho de que *un ruido blanco tiene iguales FAC y FACP*, que son:

$$\rho(k) = \varphi(k) = \begin{cases} 1, & \text{si } k = 0 \\ 0, & \text{si } k \neq 0 \end{cases}$$

Este tipo de proceso resulta de gran utilidad para construir o explicar modelos más complejos, tal como se muestra a continuación.

2.2. PROMEDIOS MÓVILES (MA)

Un **proceso de promedios móviles**, abreviado como MA(q) por sus iniciales en inglés “moving average”, es una sucesión $\{X_t\}$, con

$$X_t = \mu + Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q};$$

donde $\{Z_t\}$ es un ruido blanco con media cero y varianza σ_z^2 . Además, μ es la media del proceso MA y a los θ_i se les conoce como “los parámetros de promedios móviles e indican la intensidad de efecto del evento” (Estrada [2]). Esto es, un fenómeno que oscila alrededor de un valor μ , en donde, la oscilación y su magnitud, se explican a través de sucesos aleatorios (ruidos blancos) y las constantes que les multiplican.

Si se define $\tilde{X}_t = X_t - \mu$, (el proceso trasladado por su media), un proceso MA tiene las siguientes características:

i) $E(\tilde{X}_t) = 0,$

ii) $\text{Var}(\tilde{X}_t) = \sigma_z^2 \sum_{i=0}^q \theta_i^2$ (con $\theta_0 = 1$),

iii) $\gamma(k) = \text{Cov}(\tilde{X}_t, \tilde{X}_{t+k}) = \begin{cases} 0 & k > q \\ \sigma_z^2 \sum_{i=0}^{q-k} \theta_i \theta_{i+k} & k = 0, 1, \dots, q \end{cases}$

iv) $\rho(k) = \begin{cases} 1 & , k = 0 \\ \frac{\left[\sum_{i=0}^{q-k} \theta_i \theta_{i+k} \right]}{\left[\sum_{i=0}^q \theta_i^2 \right]} & , k = 1, \dots, q \\ 0 & , k > q \end{cases}$

2.2.1. Notas sobre los Modelos MA

Algunas observaciones y resultados que deben tenerse en cuenta para facilitar la comprensión de un modelo MA son:

1. En un proceso MA(q), la observación al tiempo t, está correlacionada únicamente con las últimas q observaciones.

2. Todo proceso MA es estacionario.
3. Un proceso MA(q) está determinado de manera única por su FAC si las raíces del polinomio

$$\theta(x) = 1 - \theta_1x - \theta_2x^2 - \dots - \theta_qx^q$$

están fuera del círculo unitario (i.e. $\theta(x_0) = 0 \Rightarrow |x_0| > 1$). ([1] pág. 34)

2.3. AUTOREGRESIVO (AR)

Un proceso $\{X_t\}$ es **autoregresivo de orden p**, denotado como AR(p), si es de la forma:

$$X_t = \mu + \phi_1X_{t-1} + \dots + \phi_pX_{t-p} + Z_t ;$$

donde μ es la media del proceso AR, $\{Z_t\}$ es un ruido blanco con media cero y varianza σ_Z^2 , y a los ϕ_i se les conoce como coeficientes autoregresivos.

Al trasladar también por su media a X_t , definiendo $\tilde{X}_t = X_t - \mu$, un proceso AR(p) cumple:

- i) $E(\tilde{X}_t) = 0$,
- ii) $\text{Var}(\tilde{X}_t) = \phi_1\gamma(1) + \dots + \phi_p\gamma(p) + \sigma_Z^2$,
- iii) $\gamma(k) = \text{Cov}(\tilde{X}_t, \tilde{X}_{t+k}) = \sigma_Z^2 \sum_{i=0}^{\infty} \phi_i\phi_{i+k}$, ($\phi_0 = 1$),
- iv) $\rho(k) = \phi_1\rho(k-1) + \dots + \phi_p\rho(k-p)$, $k \geq p$.

2.3.1. Notas sobre los Modelos AR

Se debe notar la similitud de un modelo AR con uno de regresión lineal: basta con identificar como variables explicativas a los últimos p-valores observados del proceso, y como el error a Z_t (de lo anterior se desprende el nombre de autoregresivo).

Por otra parte, para quienes conocen la teoría básica de procesos estocásticos, resulta fácil observar que un proceso AR(1) es un *proceso de Markov*.

Otros apuntes importantes de un proceso AR son:

1. Un proceso AR(p) puede describirse con sólo las últimas p-observaciones del mismo proceso.

2. Un proceso AR(p) es estacionario si las raíces de la ecuación

$$\phi(x) = 1 - \phi_1x - \phi_2x^2 - \dots - \phi_px^p$$

están fuera del círculo unitario (i.e. $\phi(x_0) = 0 \Rightarrow |x_0| > 1$).

3. Todo proceso AR está determinado de manera única por su FAC.

2.4. ARMA

Un modelo ARMA(p,q) es la combinación de un modelo AR(p) y un MA(q). Al igual que en los modelos explicados anteriormente, generalmente se utiliza el proceso trasladado por su media ($\tilde{X}_t = X_t - \mu$), cuya forma es:

$$\tilde{X}_t = \phi_1 \tilde{X}_{t-1} + \phi_2 \tilde{X}_{t-2} + \dots + \phi_p \tilde{X}_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \dots + \theta_q Z_{t-q}.$$

Con respecto a sus momentos y sobre todo a su FAC, su fórmula resulta bastante más complicada al mezclarse ya dos modelos; sin embargo, debe recordarse que existe paquetería que les estima.

2.4.1. Notas sobre los Modelos ARMA

La gran trascendencia de un proceso ARMA radica en el hecho de que *toda serie estacionaria puede ser descrita por un modelo ARMA*.

Otras observaciones de importancia, derivadas de la mezcla de los modelos son:

1. Un modelo ARMA(p,q) es la combinación de un modelo AR(p) y un MA(q).
2. Un proceso ARMA(p,q) es estacionario si las raíces de $\phi(x)$ están fuera del círculo unitario.
3. Todo proceso ARMA(p,q) está determinado de manera única por su FAC si las raíces de $\theta(x)$ están fuera del círculo unitario.

2.5. ARIMA

Hasta ahora se han descrito modelos para series de tiempo estacionarias, sin embargo, en la práctica las series generalmente no lo son. Es en este momento que surge la opción de los modelos ARIMA, en los cuales, se transforman los datos originales para obtener una serie estacionaria, y así, poder ajustar un modelo ARMA.

2.5.1. Operadores Diferencia (∇) y Retraso (B) para la Transformación de Series No-Estacionarias

La *diferenciación* de una serie es la aplicación a los datos de una transformación lineal definida como:

$$\nabla: \{\text{Series de Tiempo}\} \rightarrow \{\text{Series de Tiempo}\}$$

tal que

$$\nabla X_t = X_t - X_{t-1} .$$

En palabras, es la *obtención de una nueva serie de datos a partir de los originales, la cual se obtiene como la diferencia de cada dato con su dato predecesor o anterior*). Así, por ejemplo,

$$\nabla^2 X_t = \nabla(\nabla(X_t)) = \nabla(X_t - X_{t-1}) = X_t - X_{t-1} - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2} .$$

Por otra parte, el operador de retraso B , se define como la transformación lineal

$$B: \{\text{Series de Tiempo}\} \rightarrow \{\text{Series de Tiempo}\}$$

con

$$BX_t = X_{t-1} .$$

(Debe notarse que $\nabla = \text{Id} - B$, con Id la transformación identidad.)

2.5.2. Forma de un Modelo ARIMA

Si se denota como $W_t = \nabla^d X_t$ (y por lo tanto $W_t = (\text{Id} - B)^d X_t$), y se ajusta ahora un modelo ARMA(p,q) para W_t , entonces el modelo ARIMA(p,d,q) para X_t tiene la forma:

$$W_t = \phi_1 W_{t-1} + \dots + \phi_p W_{t-p} + Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q} ;$$

el cual, se puede reescribir usando los polinomios ϕ y θ (definidos en las secciones 2.3 y 2.4), así como el operador de retraso, de la siguiente manera:

$$\phi(B)W_t = \theta(B)Z_t .$$

2.5.3. Notas sobre los Modelos ARIMA

En resumen, *un modelo ARIMA(p,d,q) es un modelo ARMA(p,q) aplicado a una nueva serie diferenciada d-veces.*

Las condiciones de estacionariedad y de FAC determinada de forma única se cumplen bajo las mismas condiciones: estacionariedad si las raíces de $\phi(x)$ son de tamaño mayor a 1; y FAC única si $\theta(x)$ tiene todas sus raíces de magnitud mayor a 1.

2.5.4. ARIMA para Series Periódicas

En ocasiones el investigador nota un comportamiento periódico en la serie, esto es, observa que $X_t \approx X_{t-s}$ (por ejemplo, que cada $s = 12$ datos, se presentan registros similares).

El tamaño de “s” se puede estimar visualmente, sin embargo, debe observarse que su valor está relacionado con la frecuencia en que los datos fueron registrados (mes, día, semana, etc.), ya que al cambiar esa frecuencia, cambia el periodo.

A las series periódicas se les puede representar con un modelo ARIMA: simplemente eliminando la periodicidad con diferencias de orden “s” (i.e. **diferencias periódicas** definidas como $\nabla_s X_t = X_t - X_{t-s}$). Es importante resaltar que la nueva serie tendrá s datos menos (los del primer periodo se pierden), por lo que debe tenerse cuidado de no tomar “s” demasiado grande (en general de no sobrediferenciar), o de lo contrario, hacer lo posible por contar con una mayor cantidad de datos.

Es importante señalar que la aplicación de diferencias periódicas no es el único camino para eliminar periodicidad en una serie, pero que en la opinión particular de quien redacta este trabajo, es bastante intuitiva y suficiente para abordar por primera vez un análisis de series de tiempo. Si el lector se encuentra interesado en conocer otras técnicas de eliminación de periodicidad, puede consultar Brockwell & Davis [17].

2.5.5. Predicción con ARIMA

Una vez que ya se ha aceptado que un modelo ARIMA representa aceptablemente a un fenómeno, se puede cumplir otro gran objetivo de las técnicas de series de tiempo (y de la estadística en general), la *predicción o proyección de los valores de la serie*; i.e., el comportamiento futuro del fenómeno.

Para **predecir puntualmente**, esto es, para dar un único número que aproxime el valor futuro de interés, “simplemente” se despeja X_t del modelo ARIMA y, posteriormente, se sustituyen los valores observados de la serie que se involucren en la ecuación, así como los **residuos** que sustituyen a los Z_t (un residuo se define como el *valor ajustado menos el valor observado*, mientras que a los Z_t también se les suele llamar **choques aleatorios**).

La dependencia que existe entre la predicción y las observaciones anteriores puede expresarse como:

$$\hat{X}_t(k) = E(X_{t+k} \mid X_t, X_{t-1}, \dots, X_1),$$

donde $\hat{X}_t(k)$ es el estimado de X_{t+k} , y k es el tiempo al cual se va a predecir.

Según Guerrero [8], se determina un **intervalo de confianza para un pronóstico puntual**, utilizando la siguiente expresión:

$$I.C._{1-\alpha}(\hat{X}_t(k)) = \hat{X}_t(k) \pm \xi_{\frac{\alpha}{2}} \left[\sum_{j=0}^{k-1} \psi_j^2 \right]^{\frac{1}{2}} \sigma_Z,$$

con σ_Z la desviación estándar de los choques aleatorios, $\xi_{\frac{\alpha}{2}}$ el cuantil de la función de densidad normal estándar correspondiente a un nivel de confianza α , y ψ_j proviene del despeje de X_t en

$$\phi(B)X_t = \theta(B)Z_t,$$

de la siguiente forma:

$$X_t = [\phi(B)]^{-1}\theta(B) Z_t = \psi(B) Z_t,$$

y

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j, \quad \psi_0 = 1.$$

Para terminar este capítulo, y con la finalidad de ayudar a un lector poco acostumbrado a la notación matemática, *se recuerda que existe paquetería que calcula los coeficientes ψ_j , e incluso, los intervalos de confianza* (véase el anexo A.3).

Lo que sí es importante tener muy presente, es que *no basta con dar un valor como predicción, sino que se necesita saber el error en que regularmente se incurre*, y para eso sirve el intervalo de confianza. Además, el tamaño de este intervalo depende de los parámetros del modelo ajustado, de la varianza de los choques aleatorios y del nivel de confianza deseado, y, como casi todos los intervalos de confianza, está centrado en el valor puntual estimado.

Capítulo 3

SELECCIÓN DEL MODELO

A lo largo de este capítulo se describirá el método propuesto por Box & Jenkins [6] para la obtención de un modelo que describa apropiadamente a una serie de tiempo. Dicho método (mostrado en la figura 3.1) consta de cuatro etapas y es un proceso iterativo que, por ello, permite modificar el modelo hasta encontrar el mejor según los criterios del investigador.

Ya que los modelos ARMA se aplican sólo a series estacionarias, este capítulo se inicia describiendo las principales técnicas que permiten la transformación de una serie, de forma que, se establezca la varianza (se obtenga una nueva serie con varianza constante) y la media (se obtenga una nueva serie con media constante). Finalmente, se describen cada una de las etapas del ajuste de un modelo.

3.1. OBTENCIÓN DE UNA SERIE ESTACIONARIA

Además de la periodicidad, una media o una varianza no constante son las formas más comunes de no-estacionariedad. La periodicidad fue comentada en la sección 2.5.4, las otras dos se mencionarán a continuación. Cabe resaltar que *en el caso de que la serie presente los dos últimos tipos de comportamiento, primero debe estabilizarse la varianza y luego la media.*

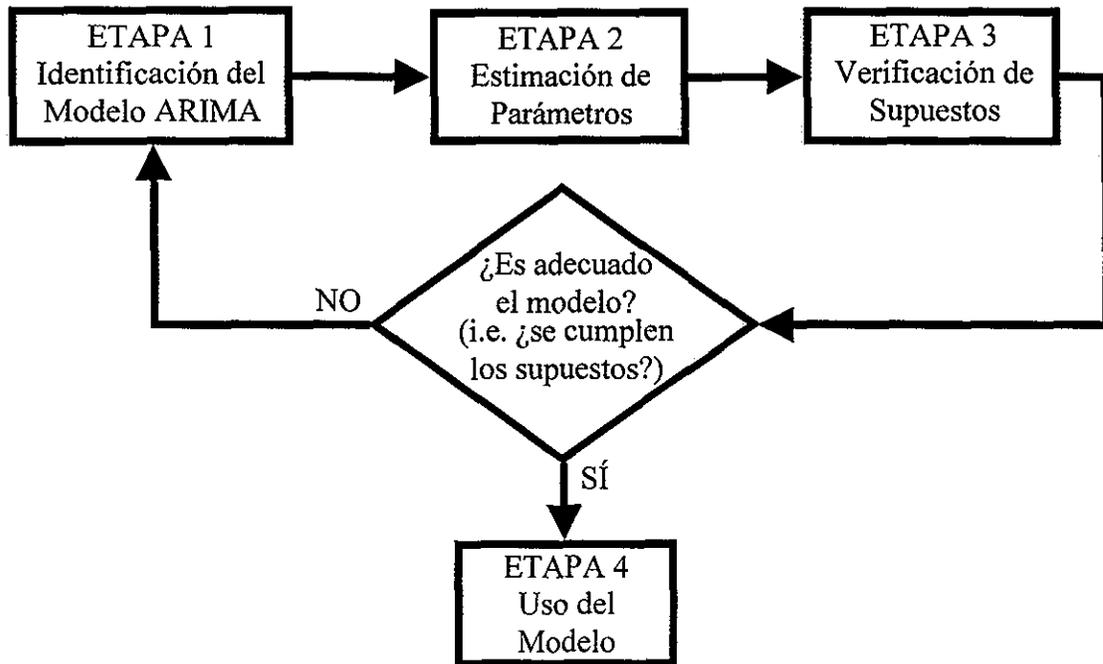


Figura 3.1. Método Box-Jenkins para la selección de un modelo.

3.1.1. Estabilización de la Varianza

Si la varianza presenta cambios a lo largo del tiempo, se puede corregir aplicando transformaciones como $\ln(x_t)$ o $\sqrt{x_t}$.

Guerrero [8], propone una forma de detectar si la varianza es constante o no; dicha técnica es también utilizada por Estrada [2] y proporciona, además, la transformación necesaria dado el caso.

De forma general, la técnica de estabilización de varianza consiste en encontrar un número real λ , de forma que

$$x_t = T(x_t) = \begin{cases} x_t^\lambda & , \lambda \neq 0 \\ \ln(x_t) & , \lambda = 0 \end{cases}$$

TESIS CON
FALLA DE ORIGEN

resulte ser una serie estacionaria. A continuación se explican los pasos para la búsqueda de λ .

Paso 1. Dividir el total de datos en H grupos con R elementos contiguos cada uno (es posible que algunos datos no pertenezcan a algún grupo). Debe resaltarse que la división será propuesta después de la inspección visual de los datos, de forma que en cada grupo la varianza parezca constante.

Paso 2. Calcular la media y desviación estándar de cada grupo; esto es, \bar{X}_h y S_h , para $h = 1, \dots, H$.

Paso 3. Sin que exista una regla (i.e. al criterio del investigador), escoger r-valores posibles para λ , todos entre -1 y 1 . Esto es, escoger $\lambda_1, \lambda_2, \dots, \lambda_r \in [-1, 1]$.

Paso 4. Evaluar un coeficiente de variación (CV) para cada λ_i , de la siguiente forma:

$$CV(\lambda_i) = \frac{\sigma_{\lambda_i}}{M(\lambda_i)},$$

donde

$$M(\lambda_i) = \frac{\sum_{h=1}^H \left(\frac{S_h}{\bar{X}_h^{1-\lambda_i}} \right)}{H}, \quad \sigma_{\lambda_i} = \sqrt{\frac{\sum_{h=1}^H \left[\left(\frac{S_h}{\bar{X}_h^{1-\lambda_i}} \right) - M(\lambda_i) \right]^2}{H-1}}$$

Paso 5. Escoger como λ a la λ_i de menor CV.

Debe notarse que si se obtiene $\lambda = 1$, es porque la serie no requiere transformarse, ya tiene su varianza constante.

Existe un método similar al anterior para la corrección de varianzas heterogéneas, y que se conoce como **transformación de Box-Cox**; el cual consiste en construir una nueva serie con la regla:

$$T(X_t) = \begin{cases} \frac{1}{\lambda} X_t^\lambda - \frac{1}{\lambda} & , \quad \lambda \neq 0 \\ \ln(X_t) & , \quad \lambda = 0 \end{cases}$$

que es básicamente la transformación propuesta por el método empírico ya descrito, excepto por la inclusión del factor $1/\lambda$ y el sumando $-1/\lambda$ (una expansión y una traslación).

En el método Box-Cox se puede buscar el mejor valor de λ por prueba y error (empíricamente), o bien, por máxima verosimilitud; y, aunque no se profundiza en el tema, debe resaltarse que el paquete estadístico S-PLUS 2000 incluye este método para transformar datos en una de sus librerías (véase el anexo A.4).

3.1.2. Estabilización de la Media

Existen varias técnicas para la detección de una media no constante, que van desde la simple inspección visual de la gráfica “tiempo vs. observaciones”, hasta el uso de las herramientas de Análisis de Regresión para probar una posible tendencia de los datos. Una digna de mención especial consiste en graficar la FAC y observar si las autocorrelaciones decaen bruscamente a cero o no (ya que *el caso negativo implica no-estacionariedad debida a la media*).

Como ya se había mencionado, si los datos presentan una media no constante, deben modificarse para poder ajustar un modelo para series estacionarias; una técnica muy utilizada para conseguirlo, es la ya explicada anteriormente: *diferenciar*.

Para determinar el grado de diferenciación requerido, además de la gráfica FAC, Anderson (citado por Guerrero [8]) resalta que *al calcular la varianza de sucesivas diferencias, ésta decrecerá hasta que la serie se torne estacionaria, pero presentará un incremento al sobrediferenciar*.

3.2. ETAPA 1: IDENTIFICACIÓN

En esta etapa se proponen los valores p , d y q que determinan a un modelo ARIMA. Como se mencionó en el capítulo introductorio, el primer paso es graficar “tiempo vs. observaciones”, y así, detectar las posibles fuentes de variación de la serie; el segundo paso es, en caso de ser necesario, transformar la serie en estacionaria; y el último, es proponer el modelo ARIMA.

Tal y como se muestra en la tabla 3.2, las gráficas de la FAC y la FACP orientan al investigador sobre el posible modelo que representa al fenómeno. Teniendo en cuenta lo mostrado en esa figura, además de la necesaria estacionariedad de la serie, en la figura 3.3 se esquematiza el proceso de identificación del modelo.

MODELO	FAC TEÓRICA	FACP TEÓRICA
AR	Tiende (decae) suavemente hacia cero.	Estadísticamente igual a cero, después de pocas (p) autocorrelaciones parciales.
MA	Estadísticamente igual a cero, después de pocas (q) autocorrelaciones.	Tiende (decae) suavemente hacia cero.
ARMA	Tiende (decae) suavemente hacia cero.	Tiende (decae) suavemente hacia cero.

Tabla 3.2. Comportamiento teórico de la FAC y la FACP.

Para decidir *cuándo los r_k o $\varphi_{k,k}$ son estadísticamente distintos a cero*, Guerrero [8] indica que para $N \geq 50$, puede usarse normalidad asintótica, y obtener las reglas:

1. En un proceso AR(p):

$$k > p \quad \& \quad |\varphi_{k,k}| \notin \left(\frac{\pm 2}{\sqrt{N-d}} \right) \quad \Rightarrow \varphi_{k,k} \text{ es estadísticamente diferente de cero.}$$

2. En un proceso MA(q):

$$k > q \quad \& \quad |r_k| \notin \pm 2 \sqrt{\frac{1}{N-d} \left(1 + 2 \sum_{j=1}^q r_j^2 \right)} \quad \Rightarrow r_k \text{ es estadísticamente diferente de cero.}$$

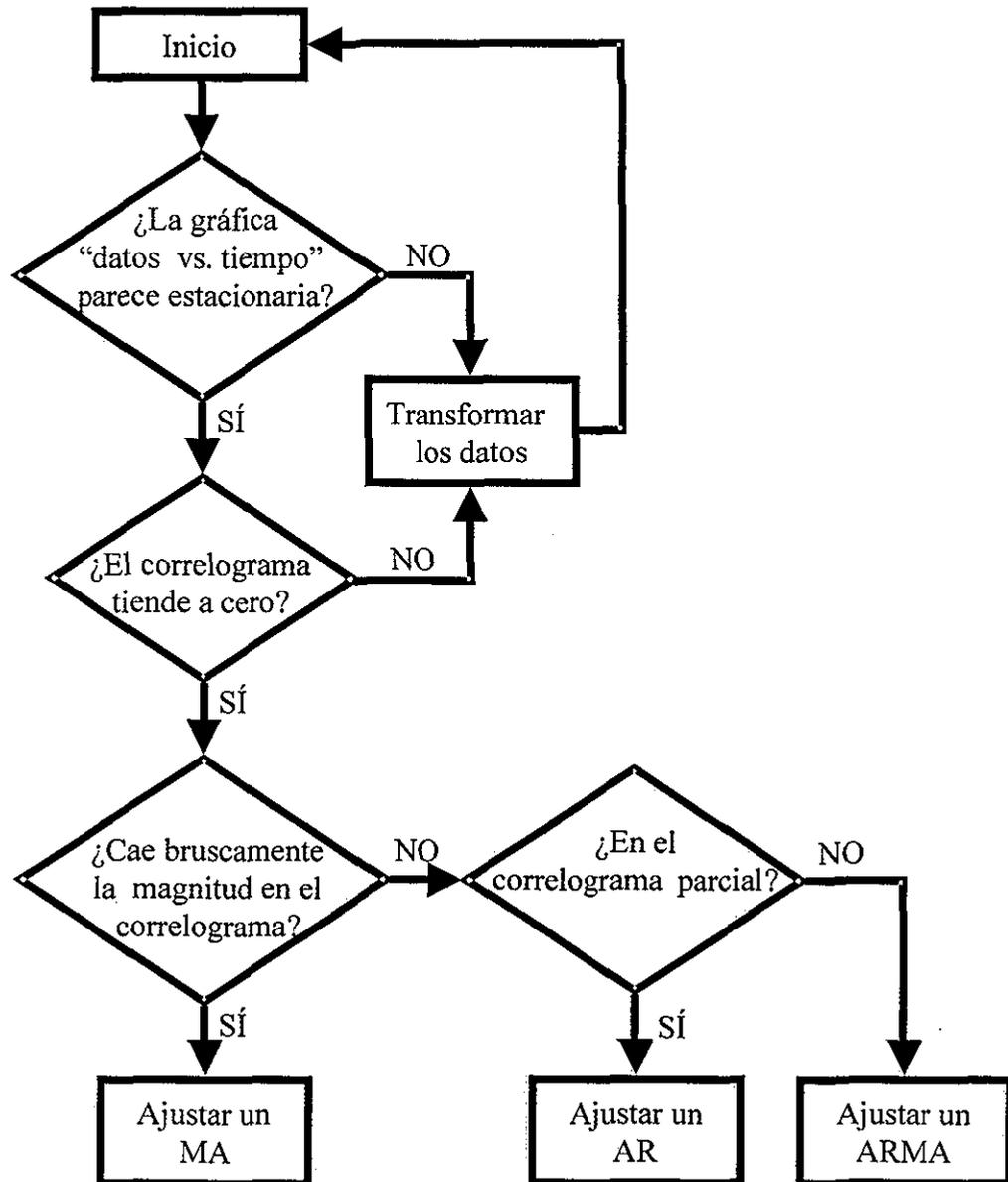


Figura 3.3. Diagrama para la identificación del modelo a ajustar.

3.3. ETAPA 2: ESTIMACIÓN

El paso siguiente a la identificación del modelo, consiste en encontrar los valores de los parámetros $\phi = (\phi_1, \dots, \phi_p)$, $\theta_0 = \mu$, $\theta = (\theta_1, \dots, \theta_q)$ y σ_z^2 ; con lo cual, se completa la identificación de un único modelo ARIMA(p,d,q).

Para la estimación de los parámetros se puede utilizar el método de Máxima Verosimilitud, que como su nombre lo indica, encuentra los parámetros de forma tal que se maximiza el valor de probabilidad de la muestra (serie) observada. Además debe recordarse que la ventaja de un estimador máximo verosímil es que resulta asintóticamente insesgado y se distribuye asintóticamente normal.

Dado un modelo ARIMA, la función de verosimilitud para los parámetros ϕ , μ , θ y σ_z^2 es:

$$L = L(X_1, \dots, X_{t-1}, X_t | \mu, \phi, \theta, \sigma_z^2) = (2\pi\sigma_z^2)^{-N/2} \exp\left(\frac{-1}{2\sigma_z^2} \sum_{t=1}^N Z_t^2\right).$$

El proceso para estimar los parámetros consiste en derivar L con respecto a μ , ϕ , θ y σ , para posteriormente, igualar a cero cada derivada y resolver un sistema de ecuaciones. ¡Bastante complicado!, por lo cual, no se profundizará en el tema; además, paquetes estadísticos como S-PLUS y SPSS se encargan de su maximización y por lo tanto, de su cálculo (véase el anexo A.3).

Chatfield [1] propone un método más fácil de entender, por su uso más frecuente en otras técnicas como la Regresión Lineal, el método de *Mínimos Cuadrados*. Este método consiste en minimizar la suma de los cuadrados de los residuos (valor ajustado – valor observado); sin embargo, se pueden obtener estimadores sesgados. Dado que la diferencia entre ambos métodos es casi siempre muy poca, se recomienda usar preferentemente el de Máxima Verosimilitud.

3.4. ETAPA 3: VERIFICACIÓN

En esta parte del proceso Box-Jenkins, partiendo de un análisis de los residuos, se diagnostica si el modelo ajustado resulta o no adecuado para el fenómeno de estudio. Debe notarse que usualmente **esta etapa no se verifica** en los trabajos estadísticos hechos en la práctica, y que, al no llevarse a cabo, **no se tiene la certeza de que el modelo describa de forma aceptable a los datos**.

Tomando en cuenta que los errores aleatorios $\{Z_t\}$ deben ser la realización de un ruido blanco, y que, para muestras grandes $Z_t \approx \hat{Z}_t$, la verificación de los supuestos se realiza con las \hat{Z}_t .

Las propiedades que deben cumplir los residuos, y que son explicadas enseguida son:

- (i) Z_t se distribuye normal para toda t .
- (ii) Z_t tiene media cero y varianza constante.
- (iii) Las Z_t 's son no correlacionadas (independientes).

Una observación importante es que en el caso de que los parámetros hayan sido estimados por el método de mínimos cuadrados, la primer propiedad puede no cumplirse.

3.4.1. Supuesto 1: Toda Z_t posee Distribución Normal

Para quienes no manejan paquetería estadística, la verificación puede visualizarse parcialmente en el histograma de residuos, en donde, en caso de existir normalidad, *se debe observar simetría* (¡pero no sólo eso!).

Otra forma de comprobación, muy intuitiva, es seguir la siguiente regla:

Si los residuos se distribuyen normalmente (y con media cero), entonces el intervalo $\left(-\xi_{1-\frac{\alpha}{2}} \hat{\sigma}_z, +\xi_{1-\frac{\alpha}{2}} \hat{\sigma}_z\right)$ contiene alrededor del $(1-\alpha)\%$ de ellos.

Utilizando la misma idea contenida en esa regla, se pueden identificar *datos atípicos*, que resultarían ser aquellos que caen fuera del intervalo $(-3\hat{\sigma}_z, +3\hat{\sigma}_z)$.

También digna de mención, resulta la gráfica de probabilidad normal (o cuantil-cuantil o Q-Q plot), en la cual se comparan los cuantiles de una distribución normal estándar con los cuantiles de los residuos estandarizados. La regla empírica es:

gráfica q-q plot “similar” a la identidad \Rightarrow normalidad.

3.4.2. Supuesto 2: Toda Z_t posee Media Cero y Varianza Constante

Recordándose que d es el grado de diferenciación, y p y q son el orden del modelo ARMA, para verificar esta propiedad en los residuos, se calculan su media y desviación estándar muestrales como

$$m(\hat{Z}) = \sum_{t=d+p+1}^N \frac{\hat{Z}_t}{(N-d-p-q)} \quad , \quad \hat{\sigma}_z = \sqrt{\sum_{t=d+p+1}^N \frac{(\hat{Z}_t - m(\hat{Z}))^2}{(N-d-p)}} \quad ;$$

y, posteriormente, aplicar una prueba de hipótesis que se resume con la regla de decisión:

$$\left| \frac{\sqrt{N-d-p-1} \ m(\hat{Z})}{\hat{\sigma}_Z} \right| \geq \xi_{1-\frac{\alpha}{2}} \Rightarrow \text{Rechazar media igual a cero con significancia } \alpha.$$

Es claro que en caso de rechazar que la media es cero, es necesario un parámetro más en el modelo (el que involucra a la media o el que elevaría el orden del modelo ARMA). Finalmente, la varianza constante puede ser visualizada en la gráfica “tiempo vs. residuos”.

3.4.3. Supuesto 3: Las Z_t 's son No-correlacionadas

Guerrero [8] describe dos técnicas para la verificación de esta propiedad, la primera consiste en una prueba de hipótesis que se resume con la siguiente regla de decisión:

$$|r_k(\hat{Z})| \geq \frac{\xi_{1-\frac{\alpha}{2}}}{\sqrt{N-d-p}} \Rightarrow \text{Rechace que } r_k = 0 \text{ con significancia } \alpha.$$

Donde

$$r_k(\hat{Z}) = \frac{\sum_{t=1}^{N-k} \hat{Z}_t \hat{Z}_{t+k}}{\sum_{t=1}^N \hat{Z}_t^2}$$

y $\xi_{1-\frac{\alpha}{2}}$ es el cuantil correspondiente en una distribución normal estándar.

La segunda técnica (debida a **BOX** y **LJUNG**) consiste en otra prueba de hipótesis, en la que se construye el estadístico

$$Q' = \frac{(N-d-p)(N-d-p+2) \sum_{k=1}^N r_k^2(\hat{Z})}{N-d-p-k};$$

el cual, si $k > 20$, se distribuye aproximadamente como χ_{k-p-q}^2 y, por lo tanto, se obtiene la regla de decisión:

$$Q' > \chi_{(k-p-q), 1-\alpha}^2 \Rightarrow \text{Rechace correlaciones iguales a cero al nivel } 1-\alpha.$$

3.5. OTRAS CONSIDERACIONES

Algo que el investigador dedicado a la estadística nombra como el “*principio de parsimonia*”, se refiere al hecho de optimizar, modelando lo mejor posible con lo menos

posible; esto es, debe observarse si todos los parámetros propuestos en el modelo son necesarios; para lo cual, se pueden construir los intervalos de confianza de cada parámetro y observar si el cero está contenido en alguno (en caso afirmativo es posible que el parámetro correspondiente sea prescindible); o, en todo caso, realizar una prueba de hipótesis formal.

Con respecto a la “*estabilidad*” del modelo (qué tanto varía al modificarse un poco los datos), debe resaltarse que ésta se pierde si los coeficientes ϕ y θ están altamente correlacionados. Para la medición de esa correlación existe una herramienta que resulta de gran utilidad en la selección del modelo, el **criterio de Akaike (AIC)**, en el cual se define

$$AIC = -2\ln L(x_{m+1}, \dots, x_N \mid x_1, \dots, x_m) + 2r \quad ,$$

el cual penaliza la logverosimilitud (logaritmo natural de la función de máxima verosimilitud) del modelo, condicionado a las primeras m observaciones y en el que r es el número de parámetros estimados. La técnica toma como ***mejor modelo a aquel que presente el AIC más pequeño***; además, utilizando el AIC, la diferencia entre dos valores es significativa si es mayor a $\chi_{s, 1-\alpha}^2$, donde s representa al número de parámetros más que un modelo tiene con respecto al otro. Ésta última sí es una ***prueba formal para decidir si un parámetro es prescindible***.

Finalmente, el lector debe tener muy claro, que *si el análisis de residuos arroja como resultado alguna violación a los supuestos, debe buscarse otro modelo* para el fenómeno; esto es, se debe de realizar siempre la etapa 3 del proceso de selección del modelo explicado a lo largo de este capítulo y esquematizado en la tabla 3.1.

3.5.1 Análisis en el Dominio de las Frecuencias

Existe otro enfoque para el estudio de series de tiempo, conocido como análisis en el dominio de las frecuencias (si se desea profundizar, puede consultarse Brockwell & Davis[9] o Chatfield[1]), en el cual, *se buscan comportamientos cíclicos que sean explicados por las funciones seno y coseno*; de forma que la serie X_t sea modelada como

$$X_t = \sum_{j=1}^k (a_j \cos(\omega_j t) + b_j \sin(\omega_j t)) + X'_t \quad ,$$

donde X'_t es un proceso estacionario, y ω_j es la **amplitud** de la componente j (en la literatura suele también definirse la **frecuencia** de la componente j como $R_j = \sqrt{a_j^2 + b_j^2}$).

Para la detección de comportamiento cíclico con esta técnica, es necesario introducir dos nuevos conceptos: la **función de densidad espectral** y el **periodograma** de una serie de tiempo.

La **función de densidad espectral** (o espectro) para una serie de tiempo, *indica la contribución a la varianza de X_t por una frecuencia ω* , y se define como

$$f(\omega) = \frac{1}{2\pi} \sum_{-\infty}^{\infty} \gamma(k) e^{-ik\omega} \quad , \omega \in (-\pi, \pi);$$

donde $i = \sqrt{-1}$ y γ es la función de autocovarianza de X_t .

Dado que en el presente trabajo se tiene una serie de números reales y no complejos, entonces $f(\omega) = \overline{f(-\omega)}$; por lo cual, sólo es necesario conocer $f(\omega)$ para $\omega \in (0, \pi)$.

El periodograma es una estimación de la función de densidad espectral, y se define como

$$I(\omega_j) = \frac{1}{N} \left| \sum_{t=1}^N X_t e^{-i\omega_j t} \right|^2,$$

donde

$$\omega_j = 2\pi j/N \quad , \quad j= 0, 1, \dots [N/2],$$

son conocidas como las **frecuencias de Fourier**, y $[N/2]$ es la parte entera de $N/2$ (para conocer su cálculo en SPSS, véase el anexo A.5).

Debe resaltarse que si el espectro del modelo ajustado es similar al periodograma basado en los datos, entonces el modelo capta la naturaleza del fenómeno; es decir, *hay una relación directa entre la similitud y modelado*. Otras observaciones importantes, que involucran a la función de densidad espectral y al periodograma, son las siguientes:

1. Un “pico” en la función de densidad espectral a la frecuencia ω_0 indica una fuerte contribución relativa a la varianza de X_t , de las frecuencias cercanas a ω_0 .
2. El espectro de un ruido blanco Z_t es constante $(\frac{\sigma_z^2}{\pi})$ en $(0, \pi)$.
3. En un proceso autoregresivo de orden mayor a dos, el espectro muestra muchos “picos”.

De los tres puntos anteriores se sigue entonces que

- 1'. Un valor grande de $I(\omega_j)$ en el periodograma, apoya la presencia de un componente sinusoidal en los datos, a la frecuencia j .
- 2'. Un periodograma uniforme (casi constante) sugiere un ruido blanco
- 3'. Un periodograma con muchos picos indica un modelo autoregresivo de orden mayor a dos.

Capítulo 4

ANÁLISIS DE LA OCUPACIÓN HOTELERA EN CANCÚN

El objetivo de este capítulo es el de modelar, para su explicación y predicción, los porcentajes de ocupación hotelera mensual observados en Cancún Quintana Roo, que claramente, pueden considerarse como una serie de tiempo $\{X_t\}$. Los datos, presentados en la tabla 4.1 , fueron obtenidos por la Secretaría de Turismo, y pueden consultarse (con todos sus decimales de exactitud) junto con otra información estadística del sector en SECTUR [5].

La selección del modelo a ajustar se hará con el método Box-Jenkins descrito en el capítulo 3; así como paquetería (EXCEL, S-PLUS y SPSS) para la realización de cálculos, estimaciones y predicciones.

4.1. TRANSFORMACIÓN A UNA SERIE ESTACIONARIA

Antes de iniciar la etapa de identificación, se construyó la gráfica “t vs X_t ” (tiempo contra porcentaje observado), mostrada en la figura 4.2. De la inspección visual de esta gráfica se concluyó que la serie parecía presentar:

- (1) comportamiento periódico,
- (2) varianza homogénea, y
- (3) media constante;

sin embargo, cada una de estas aseveraciones debieron ser constatadas.

MES	AÑO								
	1992	1993	1994	1995	1996	1997	1998	1999	2000
Enero	80.33	83.34	75.53	84.63	86.08	82.16	80.97	71.07	62.41
Febrero	84.46	86.10	79.94	88.44	89.87	89.78	84.42	77.63	80.16
Marzo	83.98	81.90	81.77	87.00	88.79	88.58	83.81	83.36	87.46
Abril	73.37	76.77	70.58	75.21	82.59	83.15	83.37	74.12	76.05
Mayo	65.51	66.19	61.89	67.39	75.90	81.18	78.07	70.38	76.04
Junio	65.60	65.27	60.08	72.28	79.33	85.78	82.58	78.19	85.82
Julio	83.04	82.02	77.26	82.09	84.78	89.17	84.95	80.78	89.48
Agosto	88.82	83.29	76.57	80.90	81.53	86.06	82.05	91.53	81.06
Septiembre	59.41	57.14	58.57	65.27	59.37	69.34	64.06	58.60	60.66
Octubre	65.56	60.53	59.36	64.95	61.13	69.82	59.10	55.21	60.12
Noviembre	72.55	63.88	62.39	72.83	72.85	75.10	66.30	58.13	64.79
Diciembre	70.68	62.04	66.04	68.89	72.07	73.20	66.41	66.08	61.92

Tabla 4.1. Ocupación hotelera mensual en Cancún.

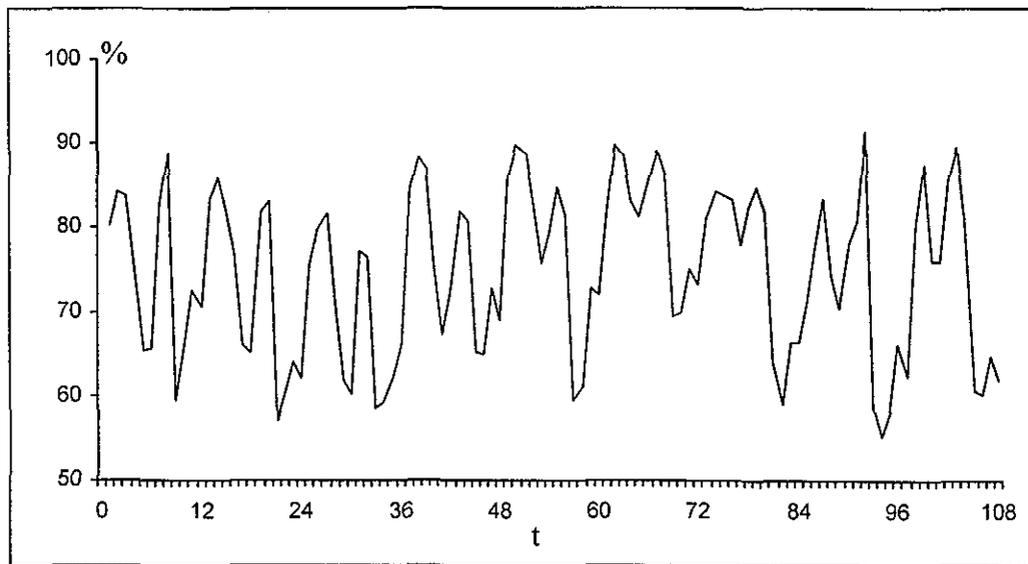


Figura 4.2. Tiempo (t) vs. ocupación hotelera en Cancún (X_t).

4.1.1. Estabilización de la Varianza

Con el propósito de obtener una serie estacionaria, el primer paso fue comprobar si la serie contaba con una varianza constante o no.

Siguiendo las ideas presentadas en la sección 3.1.1, se dividieron los 106 datos en 6 grupos, y a cada uno se le calculó su media, tal como se muestra en la tabla 4.3.

En la búsqueda de λ , para encontrar la transformación que estabilizara la varianza, se tomaron 11 valores en el intervalo $[-1,1]$, y se calculó el Coeficiente de Variación (CV), obteniéndose que el mínimo ocurría en $\lambda = -1$.

Para aproximar mejor a λ , el proceso se repitió 2 veces más, y siempre se obtuvo que el CV se minimizaba en $\lambda = -1$, (los cálculos de estas tres iteraciones en la estimación de λ se incluyen en la tabla 4.4); así, se decidió que los datos serían transformados como:

$$T(X_t) = (X_t)^{-1} = 1/X_t = Y_t$$

Para observar que sí se había logrado estabilizar la varianza, se graficó “t vs Y_t ” (figura 4.5); además, se procedió a calcular la varianza de sucesivas diferencias de Y_t , y ésta aumentaba, por lo que se dedujo que *la nueva serie ya tenía varianza aproximadamente constante*.

Finalmente, debe resaltarse que la prueba BOX-COX realizada en S-PLUS, proporcionó también $\lambda = -1$ (véase el anexo A.4).

Grupo (h)	Elementos	\bar{X}_h	S_h
1	2-18	74.85	9.28
2	20-36	68.05	9.06
3	38-54	78.11	8.51
4	56-72	77.66	9.46
5	74-90	75.88	8.24
6	92-108	71.50	12.49

Tabla 4.3. División de los datos X_t para la búsqueda de λ .

λ (iteración 1)											
	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00
$M(\lambda)$	0.0017	0.0041	0.0097	0.0230	0.0544	0.1286	0.3039	0.7187	1.6997	4.0200	9.5085
σ_λ	0.0004	0.0009	0.0021	0.0047	0.0107	0.0245	0.0559	0.1277	0.2920	0.6688	1.5348
$CV(\lambda)$	4.3743	4.5356	4.7037	4.8785	5.0594	5.2455	5.4356	5.6280	5.8207	6.0108	6.1953
λ (iteración 2)											
	-1.00	-0.98	-0.96	-0.94	-0.92	-0.90	-0.88	-0.86	-0.84	-0.82	-0.80
$M(\lambda)$	0.0017	0.0019	0.0021	0.0023	0.0025	0.0027	0.0029	0.0032	0.0035	0.0038	0.0041
σ_λ	0.0004	0.0004	0.0005	0.0005	0.0006	0.0006	0.0007	0.0007	0.0008	0.0008	0.0009
$CV(\lambda)$	4.3743	4.3901	4.4060	4.4220	4.4380	4.4541	4.4702	4.4865	4.5028	4.5191	4.5356
λ (iteración 3)											
	-1.00	-1.00	-1.00	-0.99	-0.99	-0.99	-0.99	-0.99	-0.98	-0.98	-0.98
$M(\lambda)$	0.0017	0.0018	0.0018	0.0018	0.0018	0.0018	0.0018	0.0018	0.0019	0.0019	0.0019
σ_λ	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
$CV(\lambda)$	4.3743	4.3759	4.3775	4.3791	4.3807	4.3822	4.3838	4.3854	4.3870	4.3886	4.3901

Tabla 4.4. Cálculo de Coeficientes de Variación para diferentes λ .

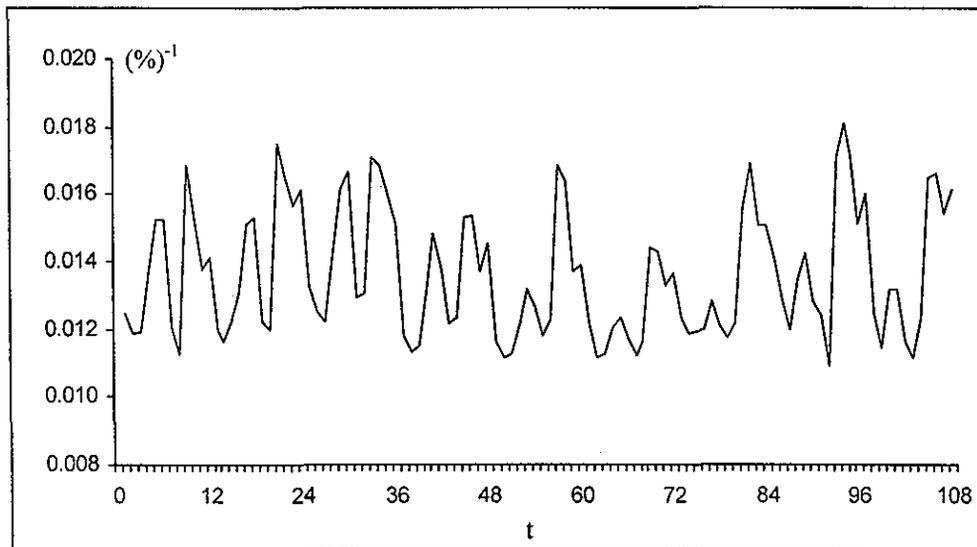


Figura 4.5. Gráfica de $Y_t = T(X_t) = X_t^{-1} = (\% \text{ de ocupación})^{-1}$.

4.1.2. Eliminación del Comportamiento Periódico

En la figura 4.5 se observa que, al igual que en la serie original, existe una probable estacionalidad en los datos. Con el objetivo de ratificar ese comportamiento de Y_t , se graficó su correlograma, incluido como figura 4.6 (véase el anexo A.2 para el cálculo de correlogramas).

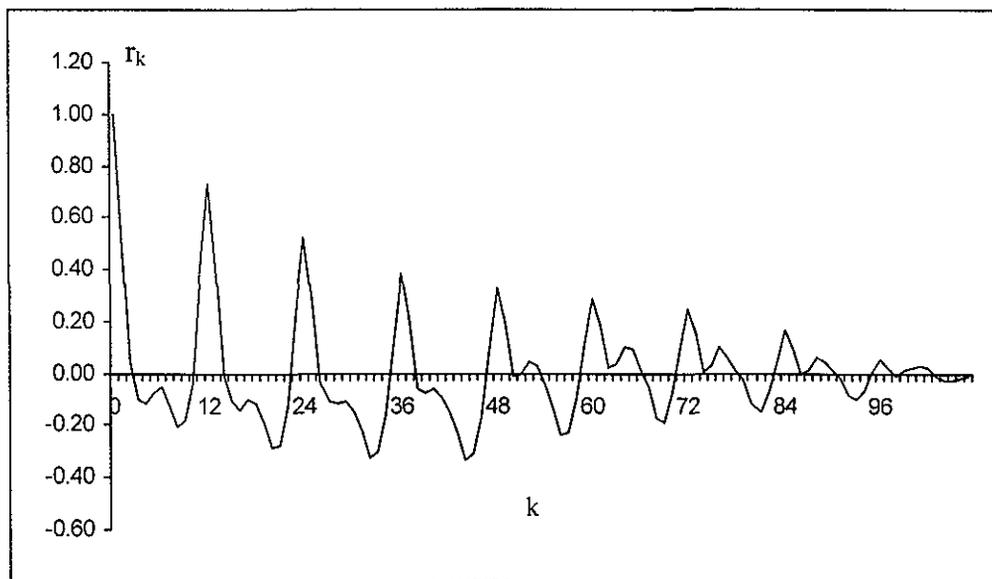


Figura 4.6. FAC de Y_t .

Con base en el correlograma de Y_t , se confirmó el *comportamiento periódico*; además, se sospechó fuertemente de un *periodo con longitud 12* (es decir, $s = 12$).

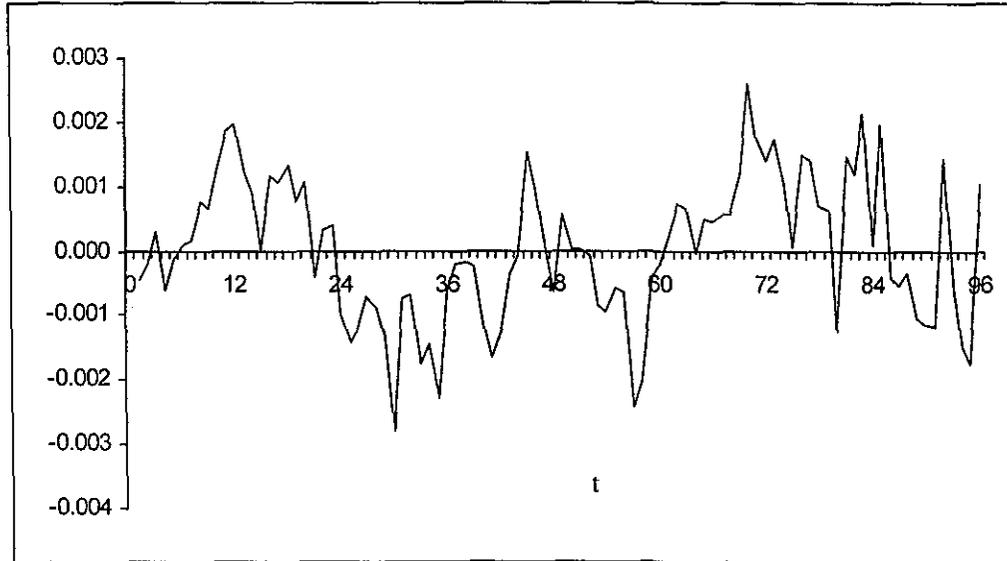


Figura 4.7. Gráfica de la nueva serie ($W_t = Y_t - Y_{t-12}$).

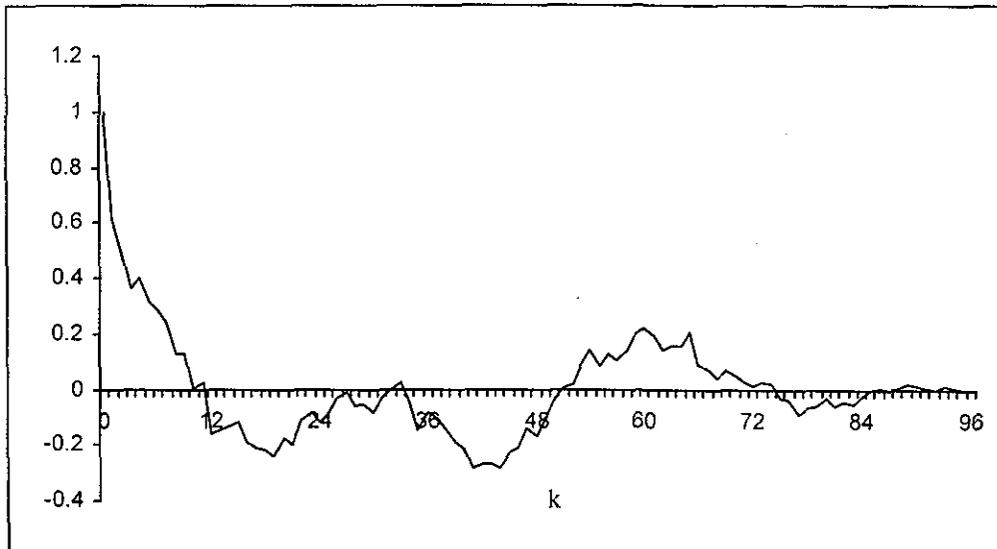


Figura 4.8. FAC de W_t .

Un experto en turismo habría dicho: “es lógico”; y sí, es claro que en cada intervalo de 12 meses hay las mismas etapas vacacionales, las mismas temporadas climáticas con sus respectivos huracanes, días calurosos, etc. También es claro que en cada una de esas épocas vacacionales (p.e. en verano), se incrementa el número de visitantes y por lo tanto el de habitaciones ocupadas.

Como ya se mencionó en el capítulo 2.5.4, existen distintas formas de eliminar la periodicidad, pero para este trabajo, se eligió la opción de diferencias periódicas.

Así, ayudado de la gráfica “t vs. Y_t ”, por el correlograma de Y_t y por el conocimiento del fenómeno, se propuso $s = 12$; esto es, se decidió transformar Y_t con la diferencia periódica ∇_{12} , obteniéndose:

TESIS CON
FALLA DE ORIGEN

$$W_t = \nabla_{12}(Y_t) = Y_t - Y_{t-12} .$$

Por último, se graficaron “t vs. W_t ” y la FAC de W_t para verificar que no se observaba ya periodicidad (figuras 4.7 y 4.8); sin embargo, se sospechó de una media no constante.

4.1.3. Eliminación de Tendencia

Para eliminar una posible tendencia, se procedió a calcular la varianza de sucesivas diferencias de W_t , y se observó que para la primer diferencia la varianza decrecía y posteriormente aumentaba (véase la tabla 4.9). Por lo que se dedujo que W_t necesitaba ser transformada como:

$$T(W_t) = \nabla(W_t) = W_t - W_{t-1} = V_t .$$

	W_t	$\nabla^1 W_t$	$\nabla^2 W_t$	$\nabla^3 W_t$	$\nabla^4 W_t$	$\nabla^5 W_t$	$\nabla^6 W_t$
Varianza	0.0000013	0.0000010	0.0000024	0.0000077	0.0000268	0.0000978	0.0003638

Tabla 4.9. Varianza de sucesivas diferencias de W_t .

En adición a esto, se ajustó un modelo de regresión simple a las V_t , y la pendiente de la recta ajustada no fue significativa (significancia de 0.925), es decir, la media de los datos era constante. Así, se concluyó que V_t era ya una serie estacionaria.

En resumen, para lograr tener una serie estacionaria se hicieron los siguientes pasos:

1. Se transformó X_t en $Y_t = (X_t)^{-1}$, para estabilizar la varianza;
2. Se transformó Y_t en $W_t = \nabla_{12}(Y_t)$, para eliminar la periodicidad;
3. Se transformó W_t en $V_t = \nabla(W_t)$, para estabilizar la media o nivel;

obteniéndose la serie mostrada en la figura 4.10.

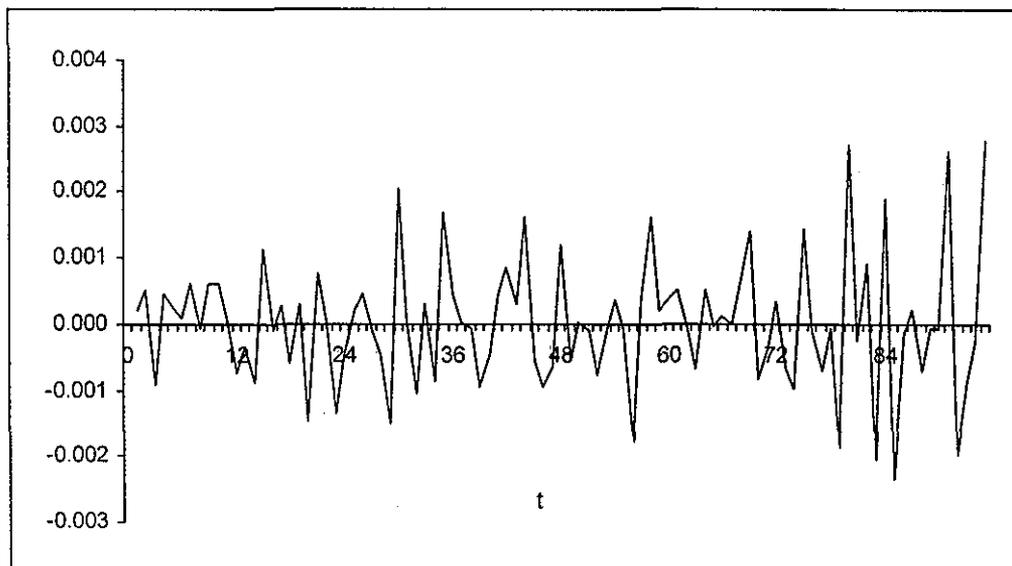


Figura 4.10. Gráfica de la serie definitiva $V_t = W_t - W_{t-1}$. “Serie ya estacionaria”.

4.2. IDENTIFICACIÓN

Una vez que se consiguió contar con una serie estacionaria, y para continuar con la notación usada en el capítulo 2, se definió la serie

$$\hat{V}_t = V_t - \hat{\mu}.$$

que se obtiene al trasladarle por su media muestral $\hat{\mu} = \hat{\theta}_0 = 0.0000154162866$.

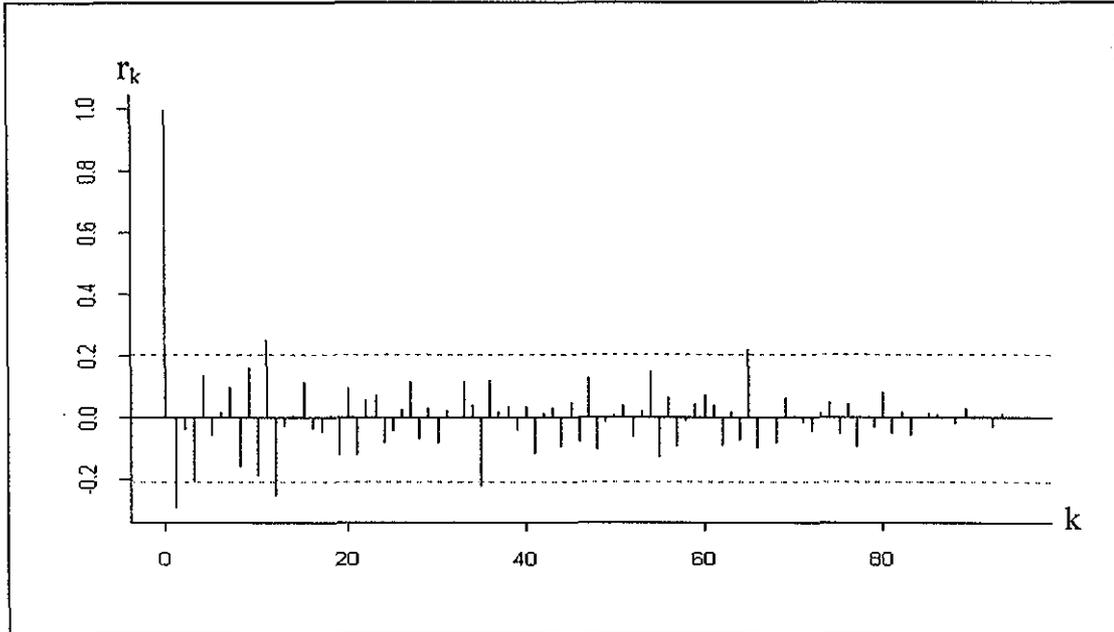


Figura 4.11. FAC de \hat{V}_t (las líneas punteadas indican qué correlaciones son estadísticamente distintas a cero).

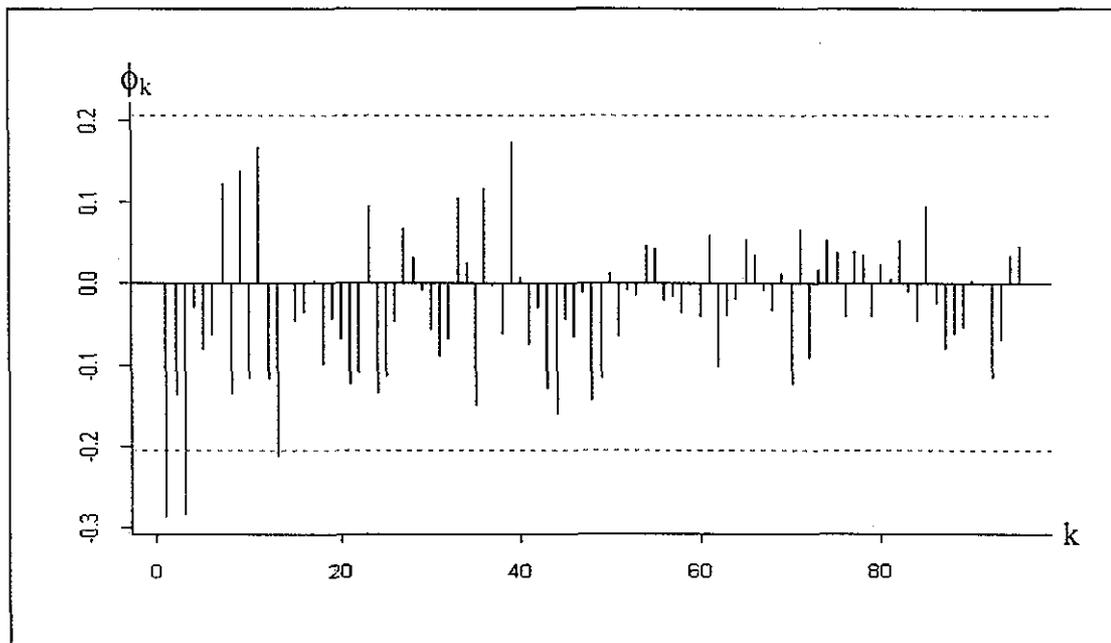


Figura 4.12. FACP de \hat{V}_t (las líneas punteadas indican qué correlaciones son estadísticamente distintas a cero).

En la figura 4.11 se muestra la FAC de \hat{V}_t , de la cual surgieron 2 posibles interpretaciones:

- ♣). *O bien la caída a cero de los coeficientes r_k es brusca después de dos correlaciones, o bien, la caída de estos coeficientes es lenta.* La primera interpretación podría ser adecuada al pensar que los coeficientes r_{11} , r_{12} , r_{35} y r_{65} fueron diferentes a cero debido al azar (ya que sólo son 4 datos de 96, equivalentes al 4.17%). La segunda surge de la idea de que en realidad, esos coeficientes todavía no son cero, de que los r_k tienden de forma suave a cero.

Por otra parte, en la figura 4.12 se muestra la FACP de \hat{V}_t , en donde se observó que:

- ♣). *Parece existir una caída brusca en la correlación parcial.* Esto porque a partir de ϕ_3 sólo hay una correlación que sale del intervalo de confianza, esto es, después de las 3 primeras correlaciones parciales, sólo una es estadísticamente diferente a cero.

En conclusión, la interpretación de los correlogramas de \hat{V}_t sugirió como posible modelo, un ARMA(3,0) o un ARMA(3,2). Sin embargo, para utilizar el criterio de Akaike y poder identificar de forma más adecuada al modelo buscado, además de esos modelos se ajustaron otros “parecidos”, tal como se muestra en la tabla 4.13 (consúltese el anexo A.3 para observar el procedimiento de ajuste utilizando SPSS).

MODELO #	ARMA(p,q)	AIC
1	(3,2)	-1,057.32
2	(3,1)	-1,058.93
3	(3,0)	-1,060.89
4	(2,2)	-1,056.81
5	(2,1)	-1,058.11
6	(2,0)	-1,054.32
7	(1,2)	-1,059.40
8	(1,1)	-1,060.13
9	(1,0)	-1,054.22
10	(0,2)	-1,059.80
11	(0,1)	-1,059.91
12	(3,3)	-1,060.21
13	(3,4)	-1,060.26
14	(4,2)	-1,057.92
15	(5,2)	-1,055.89
16	(4,0)	-1,058.88

Tabla 4.13. AIC de modelos jerárquicos cercanos al ARMA(3,2)

Se observó que el AIC mínimo se encontraba en el modelo 3; posteriormente, recordando el *sentido de parsimonia* (usar lo menos posible para modelar lo mejor posible), se probó si podía usarse un modelo con un parámetro menos, es decir, un ARMA(2,0).

Así, se realizó la prueba de hipótesis mencionada en la sección 3.5, obteniendo:

$$AIC(2,0)-AIC(3,0) = 6.57 > \chi_{1,95}^2 = 3.84 ;$$

esto es, el estadístico de prueba cayó en la región de rechazo, por lo cual, se rechazó la hipótesis nula (que en este caso es la de que el parámetro extra del modelo 3 con respecto al 6 es igual a cero). Se concluyó que entre esos dos modelos, el mejor era el 3, el ARMA(3,0).

En resumen, utilizando la inspección visual de los correlogramas, y utilizando el criterio de Akaike, *se eligió como modelo para \tilde{V}_t un ARMA(3,0)*.

4.3. ESTIMACIÓN

Los parámetros fueron estimados en SPSS (ver anexo A.3) y se muestran en la tabla 4.14 con sus correspondientes errores estándar y **significancia**. Para quienes el término resulte poco familiar, la significancia es la probabilidad de rechazar la hipótesis nula y equivocarse; que en este caso, es la probabilidad de rechazar que ϕ_i sea cero y equivocarse. Por ejemplo, si se afirma que ϕ_1 no es cero, en promedio se comete un error en 23 de cada 100 000 ocasiones.

Parámetro	Estimado	Error Estándar	Significancia
ϕ_1	$\hat{\phi}_1$		
ϕ_1	-.39668891	.10353589	.00023
ϕ_2	-.25949872	.10893322	.01927
ϕ_3	-.30493275	.10418757	.00432

Tabla 4.14. Parámetros estimados del modelo ARMA(3,0)

Se concluyó que ningún ϕ_i era estadísticamente igual a cero, por lo cual, el modelo estimado para \tilde{V}_t fue el ARMA(3,0):

$$\begin{aligned} \tilde{V}_t &= \phi_1 \tilde{V}_{t-1} + \phi_2 \tilde{V}_{t-2} + \phi_3 \tilde{V}_{t-3} + Z_t \\ &= -0.39668891 \tilde{V}_{t-1} - 0.25949872 \tilde{V}_{t-2} - 0.30493275 \tilde{V}_{t-3} + Z_t ; \end{aligned}$$

Finalmente, al sustituir

$$\tilde{V}_t = V_t - \mu ;$$

$$V_t = W_t - W_{t-1} ;$$

$$W_t = Y_t - Y_{t-12};$$

Se obtuvo el modelo ajustado a los inversos de porcentajes de ocupación en Cancún:

$$Y_t = Y_{t-12} + (1 - \phi_1 - \phi_2 - \phi_3)\mu + (\phi_1 + 1)[Y_{t-1} - Y_{t-13}] + (\phi_2 - \phi_1)[Y_{t-2} - Y_{t-14}] + (\phi_3 - \phi_2)[Y_{t-3} - Y_{t-15}] - \phi_3[Y_{t-4} - Y_{t-16}] + Z_t.$$

Posteriormente, al sustituir los valores estimados para los parámetros, el modelo encontrado fue:

$$Y_t = Y_{t-12} + 0.00003023 + (0.60331109)[Y_{t-1} - Y_{t-13}] + (0.13719019)[Y_{t-2} - Y_{t-14}] - (0.04543403)[Y_{t-3} - Y_{t-15}] + (0.30493275)[Y_{t-4} - Y_{t-16}] + Z_t.$$

Finalmente, para tener el modelo expresado en términos de la variable X_t (porcentaje de ocupación), se sustituyó $Y_t = (X_t)^{-1}$, y se obtuvo:

$$X_t = [X_{t-12}^{-1} + 0.00003023 + (0.60331109)[X_{t-1}^{-1} - X_{t-13}^{-1}] + (0.13719019)[X_{t-2}^{-1} - X_{t-14}^{-1}] - (0.04543403)[X_{t-3}^{-1} - X_{t-15}^{-1}] + (0.30493275)[X_{t-4}^{-1} - X_{t-16}^{-1}] + Z_t]^{-1}.$$

4.4. VERIFICACIÓN

En principio, los residuos \hat{Z}_t (valor real - valor estimado) observaron un *comportamiento de ruido blanco*, lo cual se percibe claramente en su FAC (figura 4.15); dado que, como se mencionó en la sección 1.6, en un ruido blanco los coeficientes de correlación son cercanos a cero.

Al buscar las raíces del polinomio $\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \phi_3 x^3$, éstas se encontraron fuera del círculo unitario, ya que:

- La única raíz real es alrededor de -1.475562 , que es de magnitud mayor a 1;
- Las raíces complejas resultaron ser aproximadamente $0.312279 \pm (1.457725)i$, cuya magnitud es 1.490799.

Por lo que el *proceso cumple con ser estacionario*.

Por otra parte, al ser un proceso AR(3), *el proceso cumple con la propiedad de invertibilidad*.

Enseguida, se procedió a analizar los residuos estandarizados (residuos divididos por su desviación estándar) para decidir si el modelo resultaba apropiado para el fenómeno.

TESIS CON
FALLA DE ORIGEN

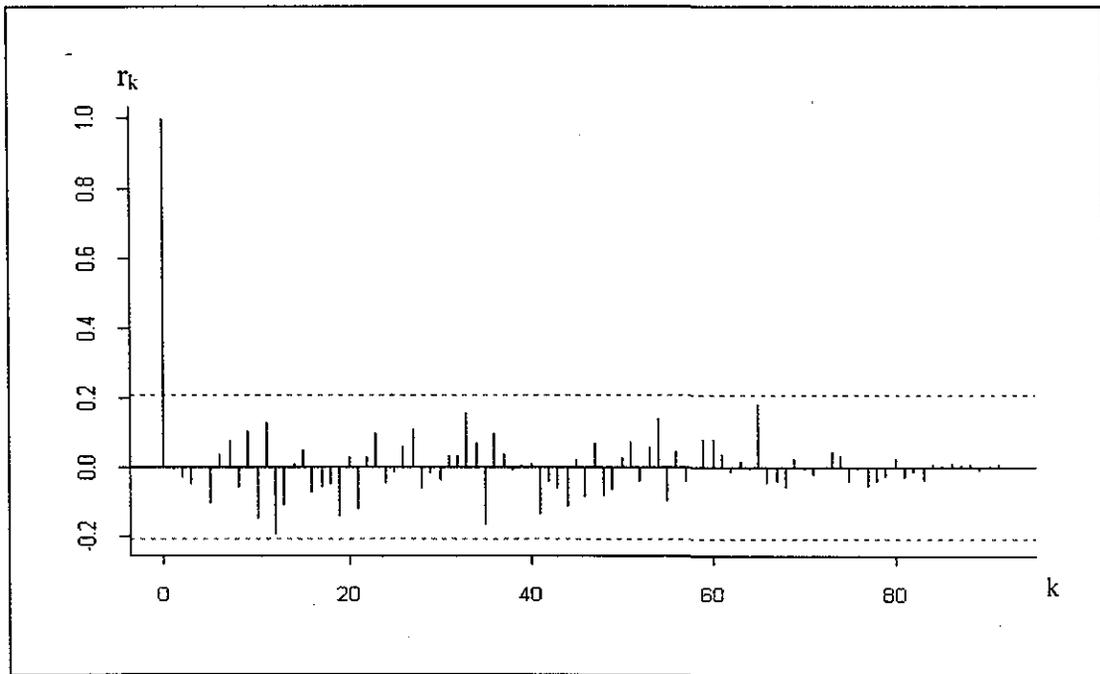


Figura 4.15. FAC de los residuos \hat{Z}_1 , que indica aleatoriedad.

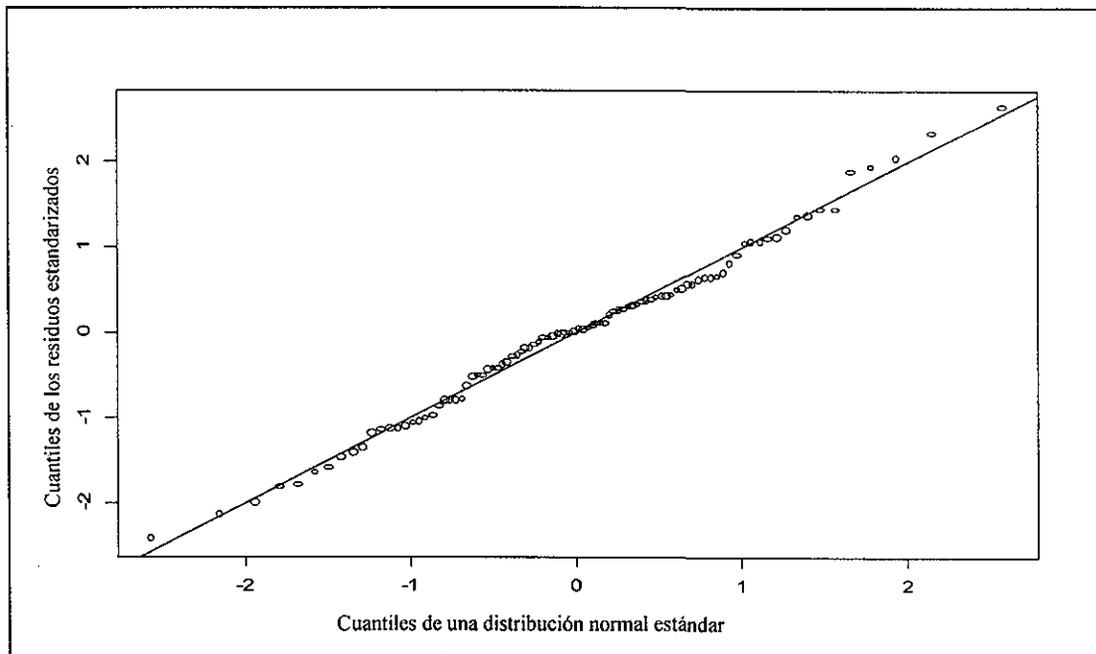


Figura 4.16. Q-Q PLOT que muestra normalidad de los residuos.

TESIS CON
FALLA DE ORIGEN

4.4.1. Normalidad de Residuos

Para comprobar si los residuos presentaban una distribución normal, utilizando S-PLUS se construyó la gráfica Q-Q PLOT (gráfica de cuantiles) de los residuos estandarizados, incluida en la figura 4.16. En ésta gráfica se aprecia que *la distribución de los residuos puede considerarse normal*.

Dada la contundencia de la gráfica q-q plot, sólo como complemento se construyó el histograma de residuos estandarizados, (figura 4.17), en el que se observa la necesaria simetría. En adición a lo anterior, es de resaltar que sólo 5 residuos estandarizados salen del intervalo ± 2 (el 5.4%); mientras que no había residuos fuera del intervalo ± 3 ; es decir, no existen residuos atípicos.

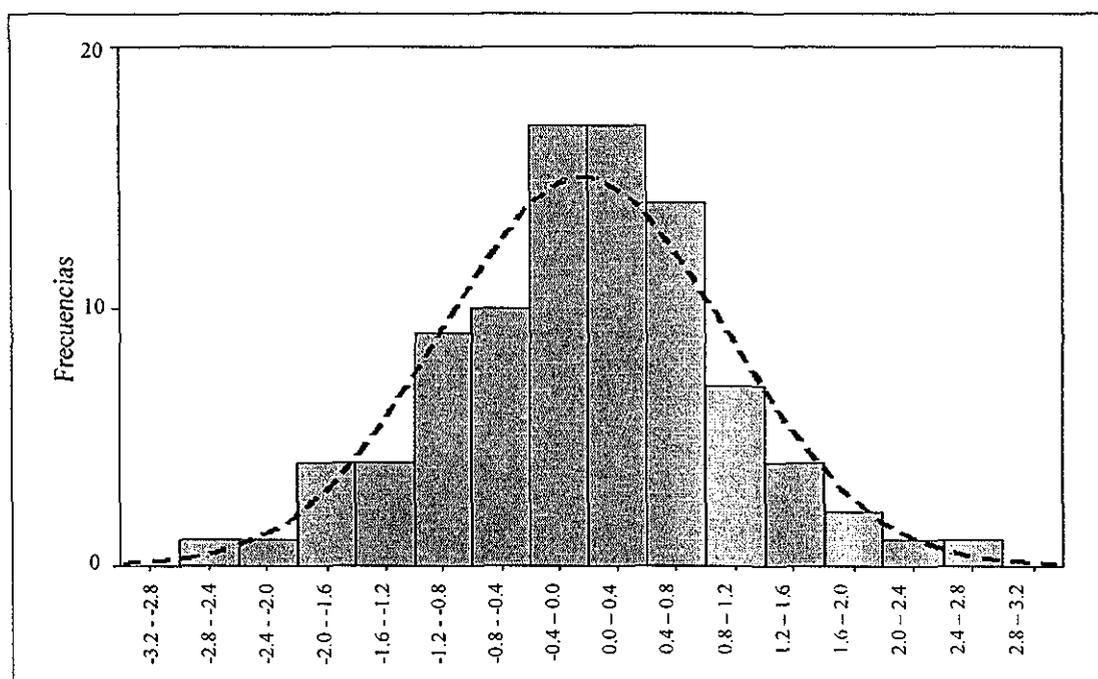


Figura 4.17. Histograma de residuos estandarizados.

4.4.2. Residuos con Media Cero y Varianza Constante

Como ya se había mencionado, los *residuos presentan un comportamiento de ruido blanco en su FAC, por lo que no hay evidencia de media o varianza no constante*. En adición a lo anterior, en la figura 4.18 se incluyó la gráfica de residuos, que de forma visual, no parece violar la característica de varianza constante.

Por otra parte, para probar que la media de los residuos era cero, se calculó el estadístico propuesto en la sección 3.4.2, obteniéndose:

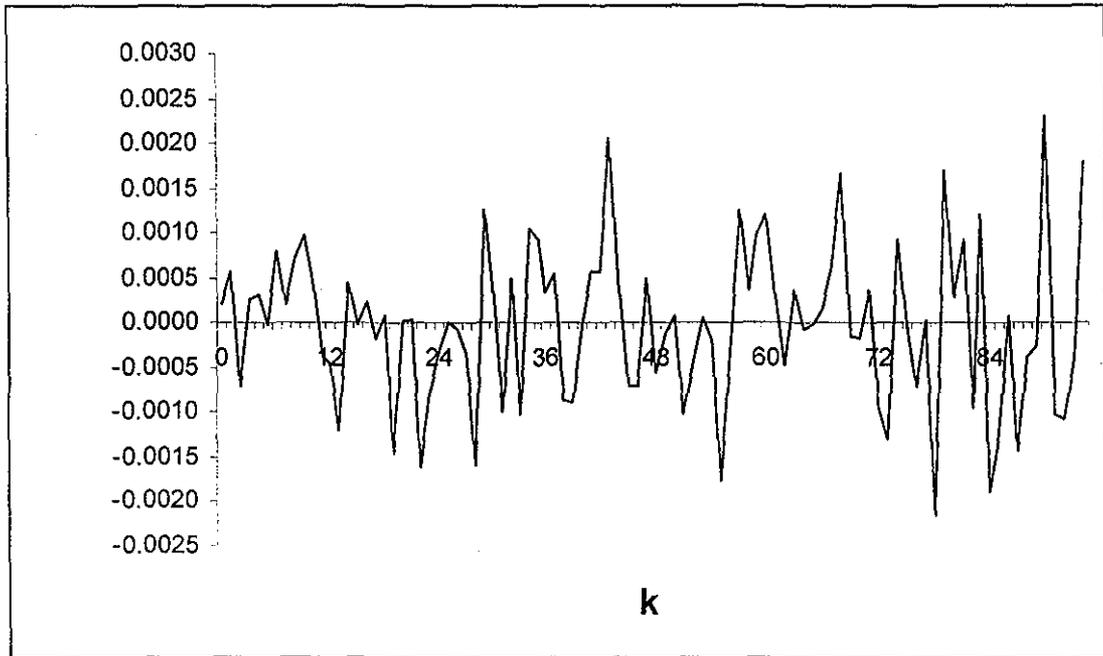


Figura 4.18. Residuos estandarizados del modelo que indican varianza constante.

$$m(\hat{Z}) = 0.000029, \quad \sigma_z = 0.000892,$$

y entonces

$$\frac{N - d - p - 1}{\sigma_z} m(\hat{Z}) = 0.305 < 1.96 = \xi_{1-\frac{\alpha}{2}},$$

con $\alpha = 0.05$. Por lo tanto, *no se rechazó media igual a cero con significancia de 5%*.

4.4.3. Residuos No-correlacionados

Además de la FAC de los residuos, que no muestra correlación entre ellos, se calculó el estadístico de Box-Ljung descrito en la sección 3.4.3 y se obtuvo:

$$Q' = 0.090004,$$

y como

$$Q' < \chi^2_{(k-p-q), 1-\alpha} = \chi^2_{92, 0.95} = 115.39,$$

no se rechazó que las correlaciones fueran iguales a cero.



4.4.4. Análisis en el Dominio de las Frecuencias

Para complementar la verificación de que el modelo ajustado resulta adecuado para el fenómeno, se utilizaron las herramientas descritas en la sección 3.5.1, útiles en la búsqueda de ciclos en una serie.

En la figura 4.19 se muestra el periodograma de la serie \hat{V}_t , el cual presenta muchos picos, característica de un modelo autoregresivo de orden mayor a dos; así, se fortaleció la propuesta de modelar con un AR(3) (consúltese el anexo A.5 para observar el cálculo del periodograma).

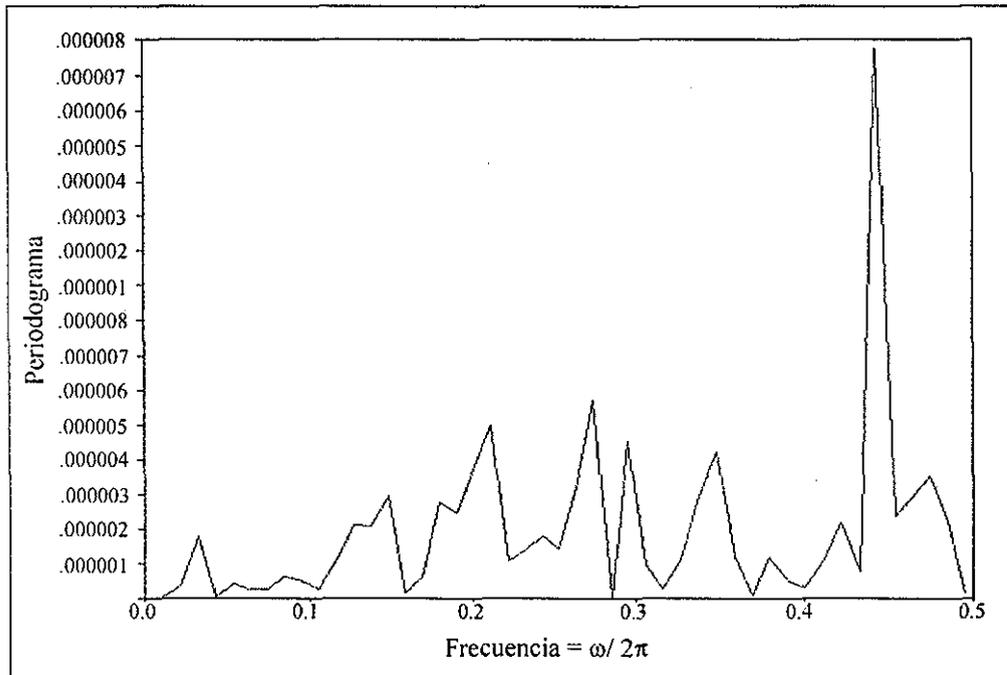


Figura 4.19. Periodograma de \hat{V}_t , (compatible con un modelo autoregresivo de orden mayor a 2).

En la figura 4.20 se calculó el periodograma de los datos ajustados, es decir, del modelo. Su similitud con la gráfica obtenida para los datos sugiere que el ajuste es apropiado.

Por último, en la figura 4.21 se obtuvo el periodograma de los residuos Z_t , en él, se observa una variación uniforme sobre todas las frecuencias (ninguna frecuencia dominante); por lo cual, se concluyó que los residuos parecían presentar un comportamiento de ruido blanco.

En resumen, *el análisis en el dominio de las frecuencias no detectó comportamiento sinusoidales, y reforzó lo obtenido por el análisis en el dominio del tiempo (modelo ARMA(3,0) para \hat{V}_t , modelado aceptable del fenómeno, y, residuos con forma de ruido blanco).*

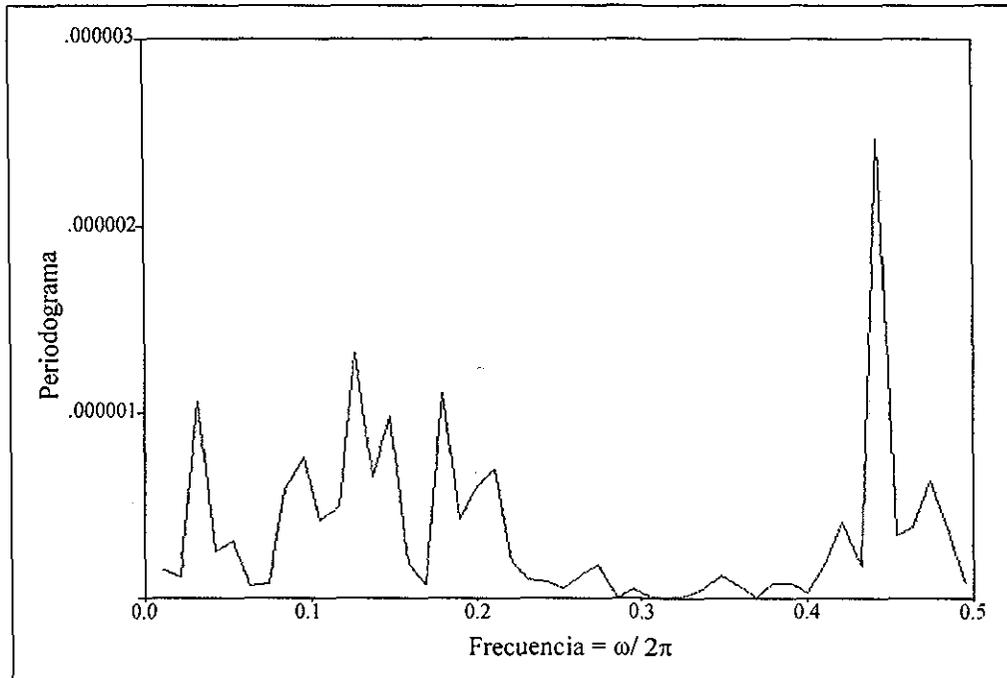


Figura 4.20. Periodograma del modelo, que indica modelado aceptable (es similar al de \hat{V}_t).

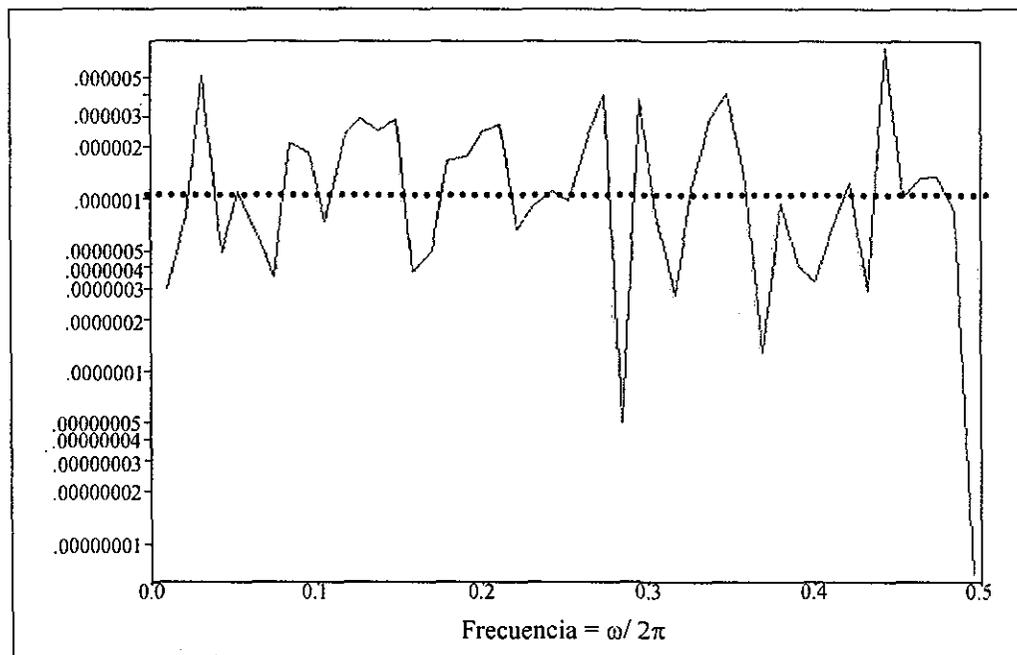


Figura 4.21. Periodograma de residuos Z_t , compatible con un ruido blanco.

4.5. PREDICCIONES

Una vez comprobados los supuestos del modelo, se procedió a realizar las proyecciones de los porcentajes de ocupación para los meses de enero, febrero y marzo de 2001; las cuales se muestran, con sus respectivos intervalos de confianza, en la tabla 4.22.

Mes	Estimación Puntual (Predicción)	Intervalo de Confianza al 95% para la Estimación
Enero de 2001	61.29	(55.35 , 68.66)
Febrero de 2001	80.33	(69.77 , 94.65)
Marzo de 2001	91.71	(78.15 , 100.0)

Tabla 4.22. Proyecciones del porcentaje de ocupación en Cancún.

Nótese el aumento en la amplitud de los intervalos, el cual es explicado porque las estimaciones para el mes de enero son con base en datos reales; mientras que a partir de febrero se involucran datos ajustados y el error en que se incurre crece.

Es oportuno señalar que, aunque *todo puede ser realizado directamente en SPSS*, las predicciones puntuales, así como los intervalos de confianza, fueron también calculados en EXCEL. Para ilustrar al lector sobre la forma en que se calculan las proyecciones e intervalos, a continuación se realizan el procedimiento para el mes de enero de 2001.

De acuerdo a lo estimado en la sección 4.3, para describir a los inversos de los porcentajes de ocupación, el modelo para pronosticar queda determinado como

$$\hat{Y}_{t+k} = Y_{t+k-12} + 0.00003023 + 0.60331109[Y_{t+k-1} - Y_{t+k-13}] + 0.13719019[Y_{t+k-2} - Y_{t+k-14}] - 0.04543403[Y_{t+k-3} - Y_{t+k-15}] + 0.30493275[Y_{t+k-4} - Y_{t+k-16}],$$

sustituyendo $k = 1$

$$\hat{Y}_{t+1} = Y_{t-11} + 0.00003023 + 0.60331109[Y_t - Y_{t-12}] + 0.13719019[Y_{t-1} - Y_{t-13}] - 0.04543403[Y_{t-2} - Y_{t-14}] + 0.30493275[Y_{t-3} - Y_{t-15}], \quad \dots(\clubsuit\clubsuit);$$

y dado que

$$\begin{aligned} Y_{t+1} &= (\text{porcentaje en enero de 2001})^{-1} &&= \text{"Valor buscado"} \\ Y_{t-11} &= (\text{porcentaje en enero de 2000})^{-1} &&= (62.4083618965020)^{-1}, \\ Y_t &= (\text{porcentaje en diciembre de 2000})^{-1} &&= (61.9178021369564)^{-1}, \\ Y_{t-12} &= (\text{porcentaje en diciembre de 1999})^{-1} &&= (66.0759460486820)^{-1}, \\ Y_{t-1} &= (\text{porcentaje en noviembre de 2000})^{-1} &&= (64.7913692539870)^{-1}, \\ Y_{t-13} &= (\text{porcentaje en noviembre de 1999})^{-1} &&= (58.1335500474062)^{-1}, \\ Y_{t-2} &= (\text{porcentaje en octubre de 2000})^{-1} &&= (60.1247085295979)^{-1}, \end{aligned}$$

$$\begin{aligned} Y_{t-14} &= (\text{porcentaje en octubre de 1999})^{-1} = (55.2096577577958)^{-1}, \\ Y_{t-3} &= (\text{porcentaje en septiembre de 2000})^{-1} = (60.6611671499163)^{-1}, \\ Y_{t-15} &= (\text{porcentaje en septiembre de 1999})^{-1} = (58.6012110375096)^{-1}; \end{aligned}$$

entonces, de ($\clubsuit\clubsuit$), la estimación puntual para enero de 2001 fue:

$$\hat{Y}_{t+1} = 0.016315,$$

y por lo tanto,

$$X_{t+1} = \text{Ocupación de enero de 2001} = (0.016315)^{-1} = \underline{\underline{61.29}}.$$

Por otra parte, para calcular el intervalo de confianza para la estimación puntual, se tiene que:

$$\begin{aligned} \psi_0^2 &= 1, && \text{(por definición),} \\ \sigma_z &= 0.000893311179842, && \text{(la desviación estándar del ruido blanco),} \\ \xi_{\frac{\alpha}{2}} &= 1.95996108232066, && \text{(cuantil de la normal estándar);} \end{aligned}$$

por lo cual,

$$\text{I.C.}_{.95}(Y_{t+1}) = \hat{Y}_{t+1} \pm \xi_{\frac{\alpha}{2}} (\psi_0^2)^{1/2} \sigma_z = (0.0145641134, 0.0180658237)$$

y finalmente, al sacar inversos,

$$\text{I.C.}_{.95}(X_{t+1}) = (55.35, 68.66),$$

que coincide con el que arrojan los paquetes estadísticos.

Capítulo 5

DISCUSIÓN FINAL

Debido al gran crecimiento de los flujos turísticos observado en el mundo durante las últimas décadas, los gobiernos y sectores comerciales y empresariales han prestado gran atención a esta industria, otorgando mayores recursos en la preparación de estrategias y programas que permitan su óptimo aprovechamiento y desarrollo. Con este auge en la industria turística y en el interés por estudiarle, las necesidades actuales y futuras de estadísticas han aumentado considerablemente, tanto en cantidad como en calidad.

Conociendo la problemática existente, inicié el presente trabajo con el propósito de contribuir al mejoramiento de la generación de datos procesados por el sector turismo, y en particular, de los porcentajes de ocupación hotelera.

A continuación, divido este capítulo final en dos partes: los resultados del modelo y las conclusiones.

5.1. RESULTADOS

El fenómeno de los porcentajes de ocupación hotelera observados en Cancún, resultó muy ilustrativo y “bondadoso” para el uso de las técnicas de series de tiempo. Fue posible el uso de todos los métodos descritos para la obtención de series estacionarias, así como la identificación clara del modelo; el cual, cumplió con todos los supuestos; y, además, las proyecciones realizadas contaron con un intervalo de confianza de tamaño aceptable.

Para conseguir una serie estacionaria se siguieron los pasos:

(1) se transformó X_t en $Y_t = (X_t)^{-1}$, para estabilizar la varianza;

- (2) se transformó Y_t en $W_t = \nabla_{12}(Y_t)$, para eliminar la periodicidad;
- (3) se transformó W_t en $V_t = \nabla(W_t)$, para eliminar posible tendencia.

Con el uso de las funciones de autocorrelación (FAC) y de autocorrelación parcial (FACP), además del criterio de Akaike, se identificó un modelo AR(3); es decir, se propuso un modelo ARMA(3,0) para \tilde{V}_t . Así, el modelo para $Y_t = (X_t)^{-1}$ fue un ARIMA(3,1,0)₁₂, el cual, al estimar los parámetros, quedó descrito como:

$$Y_t = Y_{t-12} + 0.00003023 + (0.60331109)[Y_{t-1} - Y_{t-13}] + (0.13719019)[Y_{t-2} - Y_{t-14}] - (0.04543403)[Y_{t-3} - Y_{t-15}] + (0.30493275)[Y_{t-4} - Y_{t-16}] + Z_t \dots (\clubsuit\clubsuit)$$

Se cumplieron los supuestos de normalidad de residuos, residuos con media cero y varianza constante, independencia de residuos, invertibilidad del modelo, estacionariedad y errores con forma de ruido blanco.

Con el modelo ajustado, los porcentajes no muestran tendencia, y los residuos no incluyen ciclos sinusoidales. Por otra parte, aunque la desviación estándar del ruido blanco Z_t fue 0.000892, que parece muy poca y podría ayudar a tener intervalos de confianza pequeños, al sacar inversos para obtener el estimado de los porcentajes de ocupación X_t , el tamaño de los intervalos crece. Por ejemplo, si $t+1$ corresponde a enero de 2001, se tiene:

$$I.C._{.95}(Y_{t+1}) = (0.0145641134, 0.0180658237) \quad \text{y} \quad I.C._{.95}(X_{t+1}) = (55.35, 68.66).$$

5.2. CONCLUSIONES

Inicio estas conclusiones con algunas interpretaciones del modelo.

- 1) Se puede determinar parcialmente el porcentaje de ocupación del mes t como función de: su valor el año anterior ($t-12$), los últimos 4 meses ($t-1$, $t-2$, $t-3$, $t-4$), y de esos mismos meses pero del año anterior ($t-13$, $t-14$, $t-15$, $t-16$). *Por ejemplo, septiembre de 2001 depende de septiembre de 2000, de mayo, junio, julio y agosto de 2001, y de mayo, junio, julio y agosto de 2000.*
- 2) La periodicidad en los datos se ve reflejada dentro del modelo, en el hecho de que la primer aproximación de un porcentaje es el dato observado 12 meses atrás. Esto es, el dato con mayor peso, o el que más información proporciona sobre lo que puede ocurrir en un mes t , es el mismo mes del año anterior ($t-12$). *Por ejemplo, el porcentaje con mayor peso para la predicción de septiembre de 2001, es el observado en septiembre de 2000.*
- 3) Una diferencia positiva de los meses $t-13$ y $t-1$ (es decir $X_{t-13} > X_{t-1}$) contribuye negativamente al porcentaje de ocupación del mes t . Lo anterior se sigue de que

$$\begin{aligned} X_{t-13} > X_{t-1} &\Rightarrow (X_{t-13})^{-1} < (X_{t-1})^{-1} \Rightarrow Y_{t-13} < Y_{t-1} \Rightarrow Y_{t-13} < Y_{t-1} \\ &\Rightarrow Y_{t-1} - Y_{t-13} > 0 \end{aligned}$$

Y por (♣♣), Y_t aumenta, y en consecuencia, X_t disminuye. *En palabras y con un ejemplo, si el porcentaje disminuye de agosto de 2000 a agosto de 2001, se espera que el porcentaje de septiembre de 2001 también disminuya. Lo anterior también se cumple si en lugar de utilizar el término "disminuye", se usa "aumenta".*

Siguiendo un razonamiento parecido al del inciso anterior, se tiene también que:

- 4) Una diferencia positiva en los datos de los meses $t-14$ y $t-2$ contribuye negativamente al porcentaje del mes t . *Como ejemplo, un aumento (disminución) en el porcentaje de ocupación de julio de 2000 a julio de 2001, contribuye a un aumento (disminución) en el porcentaje esperado para septiembre de 2001.*
- 5) Contrario a 3 y 4, una diferencia positiva entre los datos de los meses $t-15$ y $t-3$ contribuye positivamente al porcentaje del mes t . *Ejemplo: un aumento (disminución) en el porcentaje de ocupación de junio de 2000 a junio de 2001, contribuye a un disminución (aumento) en el porcentaje esperado para septiembre de 2001.*
- 6) De nueva cuenta, una diferencia positiva en los datos de los meses $t-16$ y $t-4$ contribuye negativamente al porcentaje del mes t . Nuevamente, *como ejemplo, un aumento (disminución) en el porcentaje de ocupación de mayo de 2000 a mayo de 2001, contribuye a un aumento (disminución) en el porcentaje esperado para septiembre de 2001*

No se observó una posible tendencia en los datos, lo cual muestra que, *aparentemente, el ritmo de construcción de nuevos hoteles en Cancún ha equilibrado al aumento en el número de turistas.*

Esta última afirmación puede constituirse como tema de una nueva investigación, donde se tomen en cuenta dos series de tiempo que interactúan: el porcentaje de ocupación y el número de habitaciones disponibles; esto es, intentar modelar el porcentaje tomando en cuenta la relación que observa con la oferta de cuartos.

Más material para nuevos trabajos, es el modelado del porcentaje de ocupación en otras localidades, ya que, claramente el fenómeno es diferente en un centro de playa que en una gran ciudad o en una ciudad fronteriza (por ejemplo, la menor ocupación en Monterrey se observa en los meses de vacaciones). Formalmente hablando, *el modelo obtenido para Cancún no debe ser aplicado a otra localidad, ya que no se pueden extrapolar sus resultados.*

También debe resaltarse que, siempre que sea posible, el modelo debe ser **reajustado**, para evitar que las predicciones incurran cada vez en un error más grande. Esto es, cuando se posean datos reales del porcentaje de ocupación, posteriores a diciembre de 2000, debe calcularse nuevamente el modelo y sus parámetros; y así, obtener mejores proyecciones.

Mes de 2001	Estimación Sector 1		Estimación Sector 2		Estimación modelo ARIMA(3,1,0) ₁₂		Valor real
	Puntual	IC 95%	Puntual	IC 95%	Puntual	IC 95%	
Enero	53.74	-	53.74	-	61.29	(55.35 , 68.66)	72.96
Febrero	82.70	-	81.48	-	80.33	(69.77 , 94.65)	80.87
Marzo	91.55	-	88.35	-	91.71	(78.15 , 100.0)	85.02

Figura 5.1. Predicciones e intervalos de confianza para los porcentajes de ocupación en Cancún.

Para ilustrar las ventajas del modelo ajustado, en la figura 5.1, incluyo el comparativo de lo obtenido con lo hecho regularmente en turismo hasta hace unos meses (las predicciones de SECTUR son hechas con dos métodos diferentes a las series de tiempo, el primero usado hasta diciembre de 1997, y el segundo utilizado, al menos, hasta julio de 2001). Aunque las predicciones son bastante parecidas, *la trascendencia del presente trabajo, es la construcción de un intervalo de confianza, es decir, se conoce el error que en promedio se está cometiendo en cada estimación.* Además, en la misma tabla se muestran los valores reales de los porcentajes de ocupación.

Debe señalarse que, para los meses de febrero y marzo, los valores reales sí se ubicaron dentro del intervalo de confianza del modelo; hecho que para el mes de enero de 2001, no sucedió. Esto puede ser interpretado como que, estadísticamente y con significancia 0.05, enero de 2001 incremento su ocupación con respecto al año anterior; posiblemente debido a algún acontecimiento como un congreso, un festival, una promoción del destino, o “algo” que no se llevó a cabo en diciembre de 2000.

Análisis como el anterior, son de gran utilidad en la evaluación de políticas turísticas, así como en la medición de sucesos económicos (como devaluaciones del peso ante el dolar), y de situaciones de inseguridad, de desastres naturales o de conflictos bélicos.

Finalmente, con respecto a los objetivos planteados al inicio de este trabajo, se logró el modelado del fenómeno con series de tiempo, consiguiendo mostrar cómo las técnicas estadísticas pueden ser utilizadas para explicar y predecir aceptablemente los porcentajes de ocupación; y con esto, dar pie a su uso en la generación de las variables procesadas en turismo. Por otra parte, mi propósito implícito de que los resultados aquí obtenidos sean revisados, adecuados, y sobre todo, utilizados en el sector público, sólo el tiempo mostrará en qué medida fue alcanzado.

ANEXOS

A.1. CUESTIONARIO HOTELERO

SECRETARIA DE TURISMO	GOBIERNO DEL ESTADO	ASOCIACIÓN DE HOTELES					
CUESTIONARIO HOTELERO		DGPT-F01					
<p>Señor gerente: El programa "cuestionario hotelero" es el instrumento para determinar la evolución de algunos indicadores de la actividad hotelera en su localidad y en nuestro país, por lo que agradecemos su colaboración en el correcto llenado y oportuna entrega de este formato. La información proporcionada está protegida por la ley de información estadística y geografía, que en su artículo 38 señala: "<i>los datos que los particulares proporcionan para fines estadísticos... serán manejados para efectos de esta ley, bajo la observancia de los principios de confiabilidad y reserva; y no podrán comunicarse, en ningún caso, en forma nominativa o individualizada, ni harán prueba ante autoridad administrativa o fiscal ni en juicio o fuera de él</i>". Diario Oficial del 30 de diciembre de 1980.</p>							
ESTABLECIMIENTO _____		DATOS DEL MES DE _____					
CATEGORÍA _____		DEL AÑO DE _____					
NÚMERO DE CUARTOS _____		LOCALIDAD _____					
		ESTADO _____					
DÍA	CUARTOS	CUARTOS OCUPADOS (A)		LLEGADA DE TURISTAS (B) No. DE PERSONAS		TURISTAS NOCHE (C) No. DE PERSONAS	
		NACIONALES	EXTRANJEROS	NACIONALES	EXTRANJEROS	NACIONALES	EXTRANJEROS
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							
29							
30							
31							
TOTAL							
(A) = CUARTOS OCUPADOS CADA DÍA, INCLUYENDO CORTESIAS. (B) = NÚMERO DE TURISTAS QUE LLEGARON CADA DÍA (ENTRADAS). (C) = CORRESPONDE AL TOTAL DE TURISTAS HOSPEDADOS AL CIERRE DEL DÍA.							

INSTRUCTIVO DE LLENADO

RECOMENDACIONES GENERALES

1. Leer cuidadosamente este instructivo antes de llenar el cuestionario.
2. Entregar el cuestionario dentro de los primeros cinco días de cada mes al funcionario de la Secretaría de Turismo Federal o Estatal designado para el efecto.

DATOS REQUERIDOS

1. DATOS GENERALES

Establecimiento. Señalar el nombre comercial del establecimiento.

Categoría. Mencionar la categoría otorgada por la Secretaría de Turismo a su establecimiento, o en su caso, la categoría comercial con la que se promociona el mismo.

Número total de cuartos. Indicar la cantidad de cuartos disponibles en el último día del mes correspondiente.

Fecha. Llenar con el mes y año que corresponden a los datos reportados.

Localización. Indicar la Ciudad y Estado en que se encuentra ubicado el establecimiento.

2. DATOS ESPECÍFICOS

Cuartos. Capturar la cantidad de habitaciones que se encuentran en servicio cada día.

Cuartos ocupados (A). Anotar el número de cuartos ocupados en el día que se reporta, incluyendo cortesías, diferenciando cuántos fueron ocupados por huéspedes nacionales y cuántos por extranjeros. Se considera turista nacional a aquel cuya dirección anotada en la tarjeta de registro pertenezca al país, en caso contrario se considera turista extranjero

Llegada de turistas (B). Anotar los turistas que llegaron (entradas) cada día reportado, diferenciando a los nacionales de los extranjeros.

Turistas noche (C). Es el total de turistas hospedados al cierre del día, clasificados según corresponda en nacionales o extranjeros.

OBSERVACIONES: Los datos específicos del cuestionario pueden ser llenados diariamente para facilitar el trabajo a la persona responsable. Para mayor información, dirigirse a:

Secretaría de Turismo
Dirección General de Política Turística,
Presidente Mazarik 172, Piso 11,
Col. Chapultepec Morales, México, D.F.

Tels. 52-50-79-48 y 52-50-74-08
Conm. 52-50-85-55
ext. 260 y 227
Fax 52-54-27-71

USO DE LA INFORMACIÓN

Los datos que usted proporcione serán utilizados para obtener los siguientes datos:

- (1) Porcentaje de ocupación (%)

$$= \frac{(\text{Cuartos ocupados por nacionales}) + (\text{Cuartos ocupados por extranjeros})}{(\text{Cuartos disponibles})} \times 100$$

- (2) Estadia o permanencia en el hotel (días)

$$= \frac{(\text{Turistas noche nacionales}) + (\text{Turistas noche extranjeros})}{(\text{Llegada de turistas nacionales}) + (\text{Llegada de turistas extranjeros})}$$

- (3) Densidad (Promedio de personas por cuarto)

$$= \frac{(\text{Turistas noche nacionales}) + (\text{Turistas noche extranjeros})}{(\text{Cuartos ocupados por nacionales}) + (\text{Cuartos ocupados por extranjeros})}$$

A.2. CÁLCULO DE LA FAC Y LA FACP CON PAQUETERÍA

Antes de dar inicio a la explicación de cómo se realizaron algunos cálculos y estimaciones con paquetería estadística, debe señalarse que la versión de S-PLUS utilizada a lo largo de este trabajo fue la 3.2 (excepto para el cálculo de la transformación Box-Cox, que fue necesaria la versión 2000); mientras que la de SPSS fue la 8.0. En adelante, siempre que se haga mención a uno de estos paquetes estadísticos, se hará referencias a estas versiones. Además, debe aclararse que el objetivo de incluir estos anexos, es el de guiar en el uso de la paquetería, y no el de proporcionar un manual sobre ella.

Debido a que fue el paquete usado en este trabajo para el cálculo de correlogramas, sólo se detalla el procedimiento en S-PLUS; sin embargo, para quien esté más familiarizado con SPSS, es oportuno señalar que este paquete también calcula la FAC y la FACP, basta con seleccionar la opción

“Graphs ↵ Time Series ↵ Autocorrelations ↵”.

Cálculo de la FAC y la FACP en S-PLUS

Para calcular la FAC o correlograma de una serie de tiempo X_t , una vez capturados los datos en S-PLUS, debe escribirse la instrucción:

“FAC1_acf(X, lag.max=k, type="correlation", plot=T)”,

y presionar “enter” (↵);

donde ‘X’ es el vector de datos X_t , ‘lag.max’ es el número de retrasos, k, que desean calcularse, ‘type’ indica si se desea la autocorrelación o la autocorrelación parcial, ‘plot’ permite graficar o no el correlograma, y en FAC1 se guarda la información del correlograma.. Es importante señalar que si se desea graficar el correlograma, debe abrirse antes una ventana de gráficos con la instrucción: “win.graph()”. Además, la información guardada es visible al escribir FAC1.

Si se desea la FACP o correlograma parcial, la instrucción es la misma sustituyendo “correlation” por “partial”.

Ejemplo

Para una mejor comprensión del procedimiento que debe seguirse, a continuación se muestra el cálculo de la FAC de \hat{V}_t , realizado en el capítulo 4.

Paso1. Se capturaron los datos en S-PLUS como la variable Vtmm (V_t menos su media), de forma que al escribir “Vtmm ↵”, en pantalla aparecieron los datos tal como se muestra en la figura A.I.

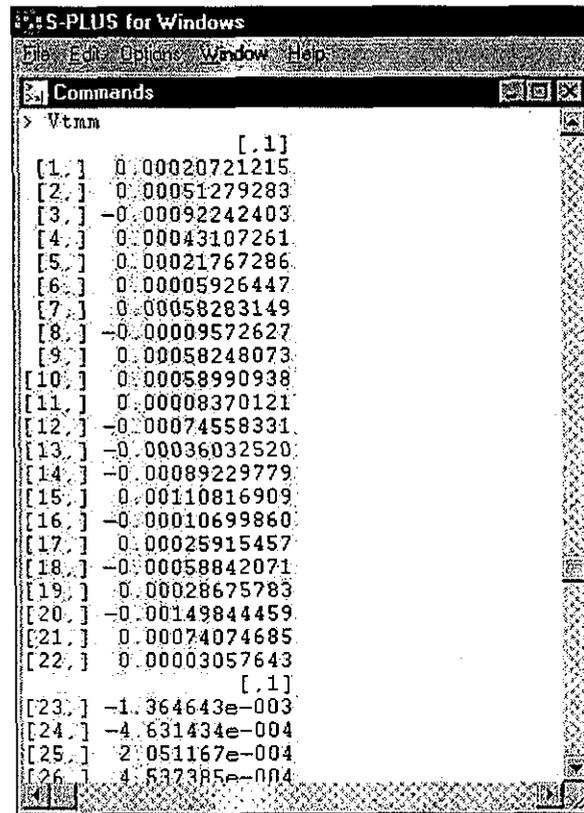


Figura A.I. La serie \hat{V}_t , capturada en S-PLUS.

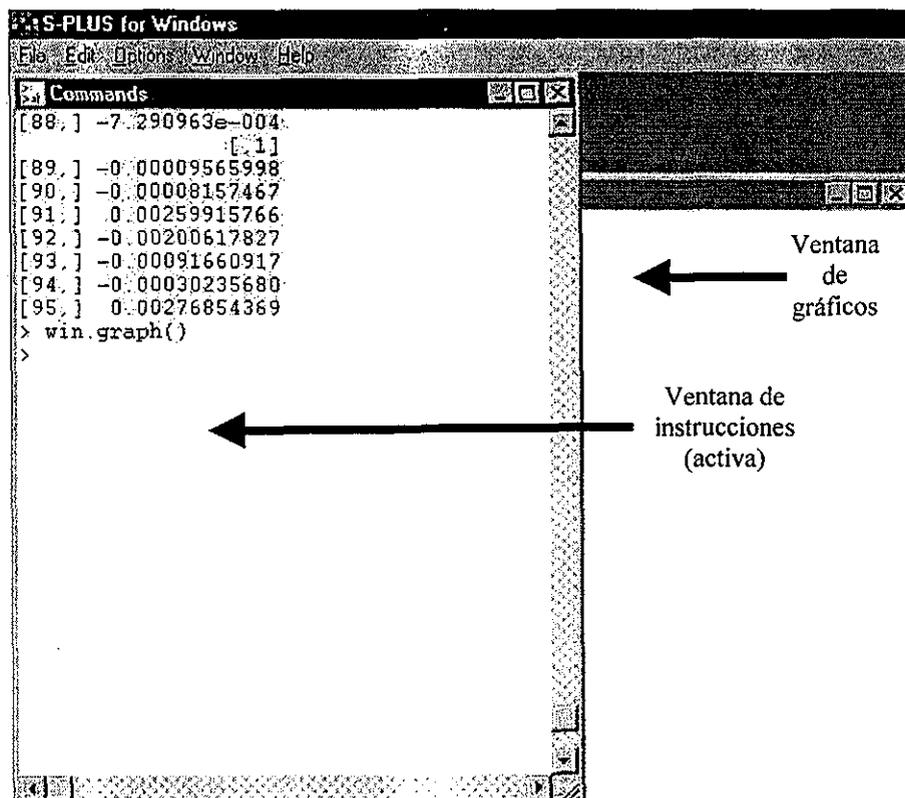


Figura A.II. Creación de una ventana de gráficos.

Paso 2. Se abrió una ventana de gráficos con la instrucción “win.graph()”, y se dejó como ventana activa la de instrucciones (figura A.II)

Paso 3. Se dio la instrucción

“FACVtmm_acf(Vtmm, lag.max=95, type="correlation", plot=T)”.

En la ventana de gráficos apareció la FAC de \hat{V}_t , con su banda de confianza al 95% (y que se incluyó en el capítulo 4 como figura 4.11).

Paso 4. (Opcional). Si se desea guardar los coeficientes de autocorrelación, puede escribirse

“FACVtmm ↵”,

estos aparecerán precedidos de la señal “\$acf” (véase la figura A.III) y se pueden copiar con la instrucción “control+insert”. Lo anterior es de utilidad si se pretende usar otro paquete para graficar el correlograma (tal como en las figuras 4.6 y 4.8).

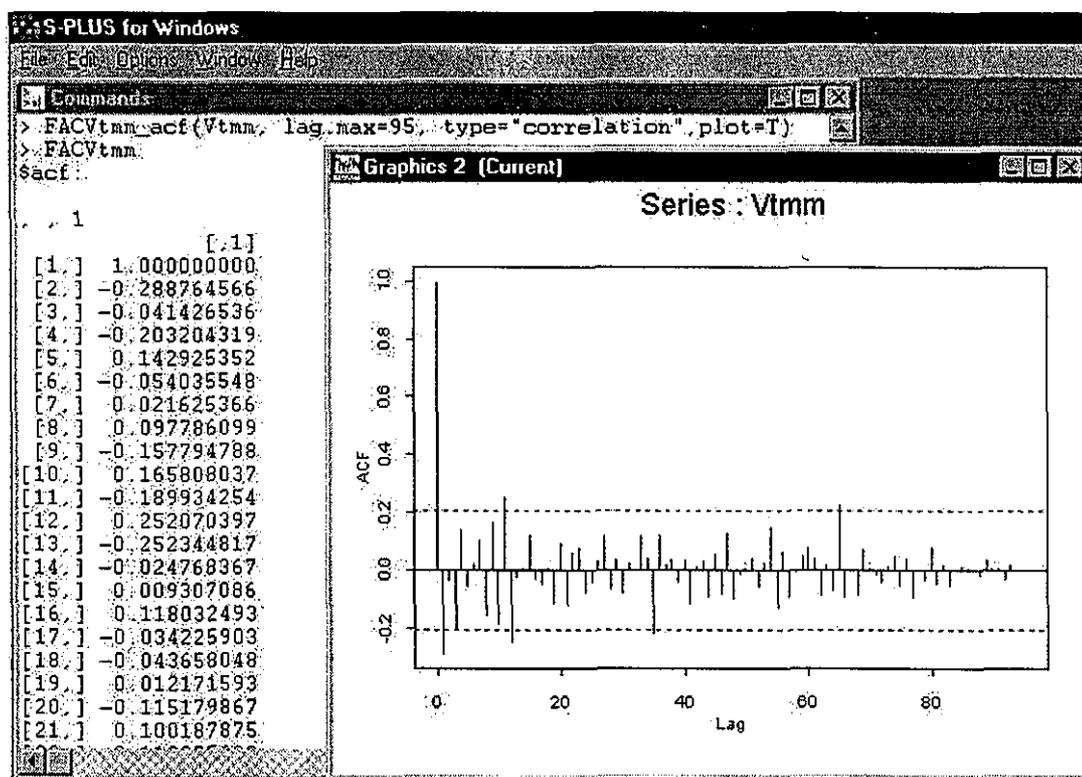


Figura A.III. Coeficientes de autocorrelación y correlograma de \hat{V}_t .

TESIS CON
FALLA DE ORIGEN

A.3. AJUSTE DE UN MODELO ARIMA EN SPSS

Aunque el ajuste de un modelo de series de tiempo puede ser realizado en S-PLUS, en esta sección se explica sólo el procedimiento seguido en SPSS, paquete que fue utilizado para ajustar el modelo propuesto en el presente trabajo.

Ajuste en SPSS

Una vez capturados los datos en SPSS, para ajustar un modelo ARIMA se selecciona la opción

“Statistics ↵ Time Series ↵ ARIMA ↵”;

se desplegará la ventana mostrada en la figura A.IV, allí se escogen, entre otras cosas:

- Dependent: Variable que depende del tiempo (X_t).
- Transform: Transformación que desea hacerse a la variable dependiente.
- Independent: Variable que, además del tiempo, se supone explica a X_t .
- p, d y q: Orden del modelo autoregresivo, grado de diferenciación y orden de modelo MA.
- Include constant in model: Opción que al activarse estima el valor de la constante del modelo (en caso contrario, supone constante igual a cero).

La opción “Save” se utiliza para crear los datos ajustados con sus intervalos de confianza, así como las predicciones del modelo, también con sus intervalos de confianza.

La opción “options” permite decidir sobre el criterio de convergencia en la estimación de los parámetros del modelo; así como el desplegado de los valores estimados (en cada iteración o sólo los finales).

Ejemplo

Con la finalidad de explicar aún más el procedimiento, a continuación se muestra el ajuste de un modelo ARMA(3,0) para \hat{V}_t , tal y como se realizó en el capítulo 4.

Paso 1. Se capturaron los datos en SPSS como la variable Vtmm (V_t menos su media), en la forma que se ilustra en la figura A.V.

Paso 2. Debido a que la serie era ya estacionaria, que ya se le había restado su media, y que se deseaba ajustar un modelo ARMA(3,0), se llenaron los campos en la ventana tal como se ilustra en la figura A.VI.

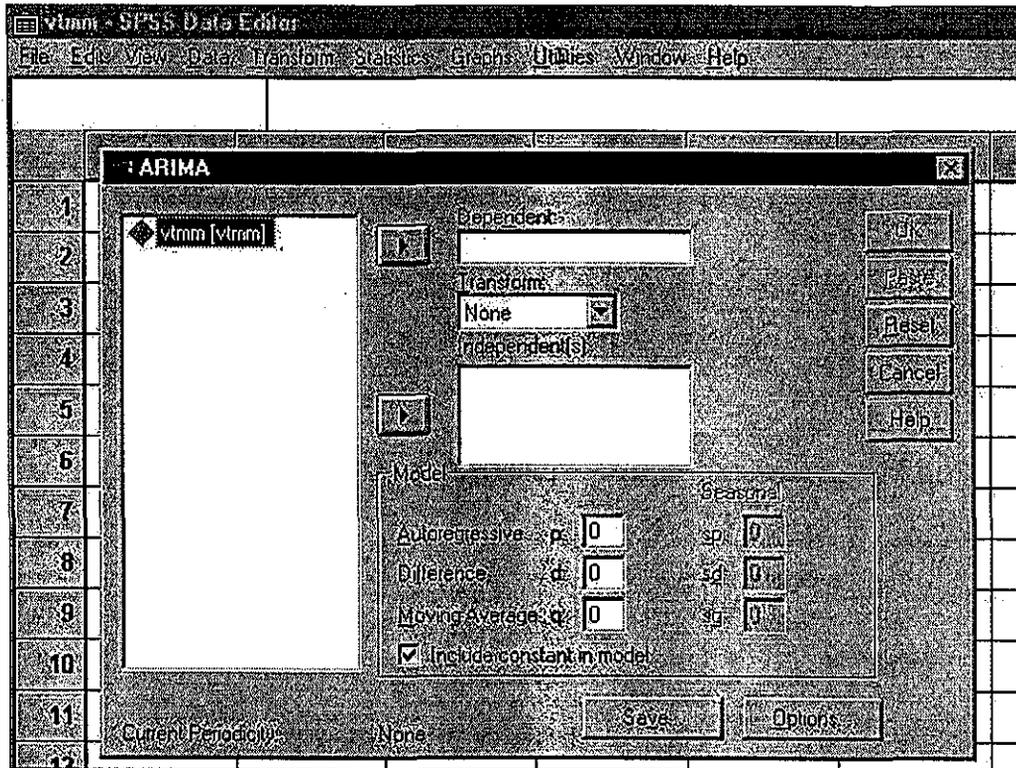


Figura A.IV. Ventana de ajuste de un modelo ARIMA en SPSS

	vtrmm	var	var	var
1	.0002072121518524			
2	.0005127928315492			
3	-.0009224240311878			
4	.0004310726109928			
5	.0002176728567043			
6	.0000592644735673			
7	.0005828314924781			
8	-.0000957262719094			
9	.0005824807250074			

Figura A.V. La serie \hat{v}_t capturada en SPSS.

TESIS CON
FALLA DE ORIGEN

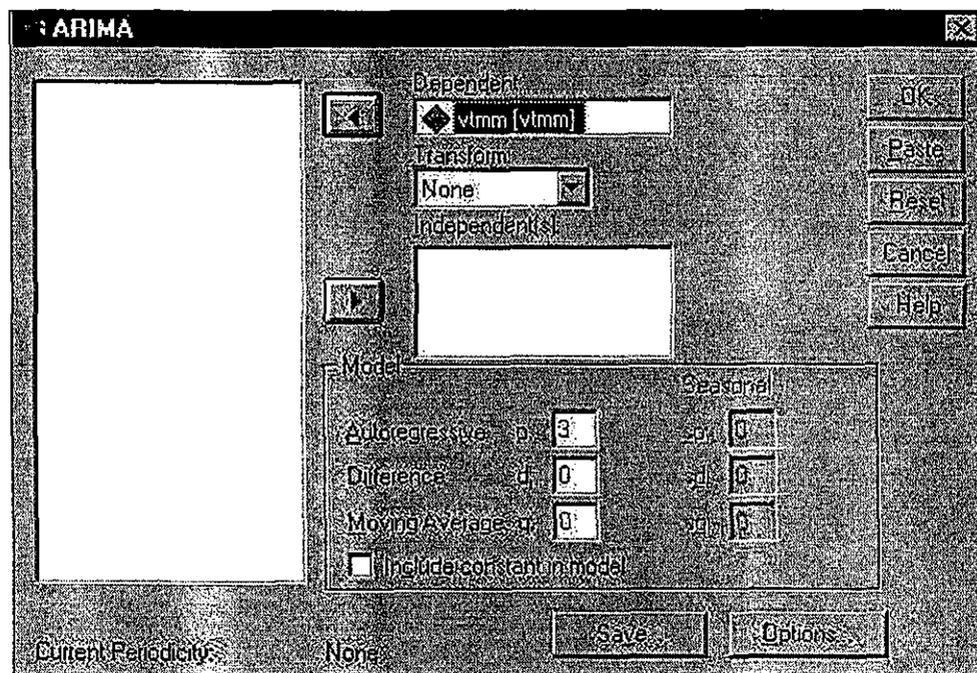


Figura A.VI. Ajustando un modelo ARMA(3,0) a la serie \hat{V}_t .

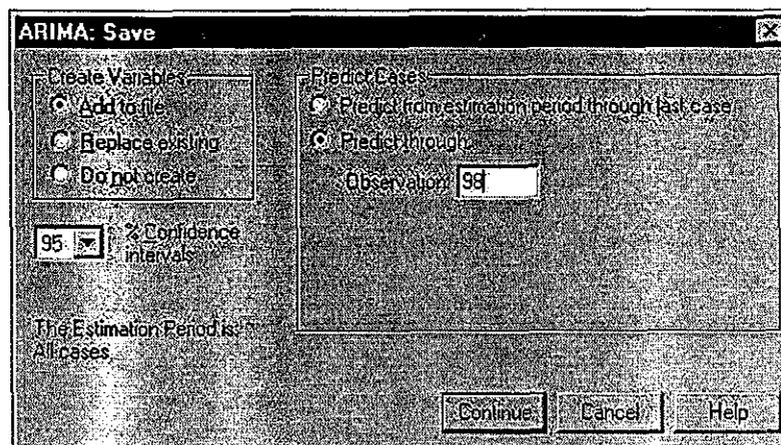


Figura A.VII. Creación de datos ajustados, predicciones e intervalos de confianza.

Paso 3. En la opción “Save” se eligió la estimación de un intervalo de confianza al 95% para cada dato ajustado o estimado; además de que se solicitaron 98 predicciones, con lo cual, el modelo ajustó 95 datos y predijo 3 (la variable \hat{V}_t constaba de 95 datos). Este paso se ilustra en la figura A.VII.

Paso 4. Se corrió el ajuste pulsando “OK”.

Paso 5. En la salida del proceso (incluida en la figura A.VIII), se obtuvieron ϕ_1 , ϕ_2 y ϕ_3 con sus intervalos de confianza, la varianza del modelo, el AIC (útil para el uso del criterio de Akaike), y las nuevas variables creadas (véase también la figura A.IX).

TESIS CON
FALLA DE ORIGEN

MODEL: MOD_1

Split group number: 1 Series length: 95
 Number of cases skipped at end because of missing values: 12
 Melard's algorithm will be used for estimation.

Conclusion of estimation phase.
 Estimation terminated at iteration number 2 because:
 Sum of squares decreased by less than .001 percent.

FINAL PARAMETERS:

Number of residuals 95
 Standard error .00089331
 Log likelihood 533.44354
 AIC -1060.8871
 SBC -1053.2255

Analysis of Variance:

	DF	Adj. Sum of Squares	Residual Variance
Residuals	92	.00007375	.00000080

Variables in the Model:

	B	SEB	T-RATIO	APPROX. PROB.
AR1	-.39668891	.10353589	-3.8314144	.00023266
AR2	-.25949872	.10893322	-2.3821816	.01926880
AR3	-.30493275	.10418757	-2.9267671	.00431499

The following new variables are being created:

Name	Label
FIT_1	Fit for VTMM from ARIMA, MOD_1 NOCON
ERR_1	Error for VTMM from ARIMA, MOD_1 NOCON
LCL_1	95% LCL for VTMM from ARIMA, MOD_1 NOCON
UCL_1	95% UCL for VTMM from ARIMA, MOD_1 NOCON
SEP_1	SE of fit for VTMM from ARIMA, MOD_1 NOCON

Figura A.VIII. Salida en SPSS del ajuste de un modelo ARMA(3,0) a la serie \hat{V}_t .

Para seguir con la notación utilizada a lo largo del trabajo, debe leerse:

AR1 = ϕ_1 ;
 AR2 = ϕ_2 ;
 AR3 = ϕ_3 ;

Estándar

error = σ_z^2 ;

Fit = Valor ajustado (o proyectado);

ERR = Residuo = Z_t ;

LCL = Límite inferior del intervalo de confianza;

UCL = Límite superior del intervalo de confianza;

SEP = Error estándar de la predicción.

Finalmente, si se activa la ventana con los datos, se observa que aparecen las nuevas variables, mencionadas en la salida del ajuste, así como las 3 predicciones solicitadas (figura A.IX).

	vtm	fit_1	err_1	lcl_1	ucl_1	sep_1
86	-.0001700670	.00108677147	-.0012568384	-.0006874216	.00266096458	.00089
87	.00020274723	.00011617718	.00008667005	-.0016580159	.00189037029	.00089
88	-.0007290963	.00068805518	-.0014171515	-.0010861379	.00246224829	.00089
89	-.0000956600	.00028847076	-.0003841307	-.0014857224	.00206266387	.00089
90	-.0000815747	.00016532253	-.0002468972	-.0016088706	.00193951564	.00089
91	.00259915766	.00027950875	.00231964891	-.0014946844	.00205370186	.00089
92	-.0020061783	-.0009807186	-.0010254596	-.0027549118	.00079347447	.00089
93	-.0009166092	.00014622539	-.0010628346	-.0016279677	.00192041850	.00089
94	-.0003023568	.00009164109	-.0003939979	-.0016825520	.00186583421	.00089
95	.00276854369	.00096954995	.00179899373	-.0008046432	.00274374307	.00089
96		-.0007402852		-.0025144783	.00103390788	.00089
97		-.0003325721		-.0022412626	.00157611841	.00096
98		-.0005201889		-.0024374621	.00139708431	.00097
99						
100						
101						
102						
103						

Figura A.IX. Variables creadas en SPSS al ajustar un modelo ARMA(3,0) a \hat{V}_t .

TESIS CON FALLA DE ORIGEN

A.4 MÉTODO BOX-COX EN S-PLUS

Para obtener el valor de λ de la transformación Box-Cox, necesario para la estabilización de la varianza de una serie, una vez capturados los datos en S-PLUS 2000 (como vector columna al igual que en la figura A.I), debe escribirse

```
“library (MASS)” ↵ “add Mass Menus()” ↵ ,
```

esto para poder habilitar la librería que contiene la opción. Posteriormente, puede pedirse la ayuda del paquete, escribiendo

```
“help(box-cox)” ↵ ,
```

y seguir las instrucciones.

ESTA TESIS NO SALI
DE LA BIBLIOTECA

A.5 CÁLCULO DE PERIODOGRAMAS EN SPSS

Una vez que se cuenta con la información en SPSS, se escoge la opción

“Graphs ┆ Time Series ┆ Spectral ┆”

Se activa una ventana como la mostrada en la figura A.XI, y en la cual se debe seleccionar, entre otras cosas:

- Variable(s): Serie o series de datos a los que se les pretende calcular el periodograma.
- Spectral Window: Método utilizado para suavizar el periodograma, y así, estimar la densidad espectral.
- Center variables: Opción que permite quitar o no la media a los datos, antes de calcularles el periodograma.
- Plot: Permite seleccionar el tipo de gráfica, periodograma o espectro (estimado).
- By frequency/ By period: Tipo de escala usada en el eje x de la gráfica.

Ya que fue completada esta ventana, se presiona “OK” , y en la salida del proceso aparece lo solicitado.

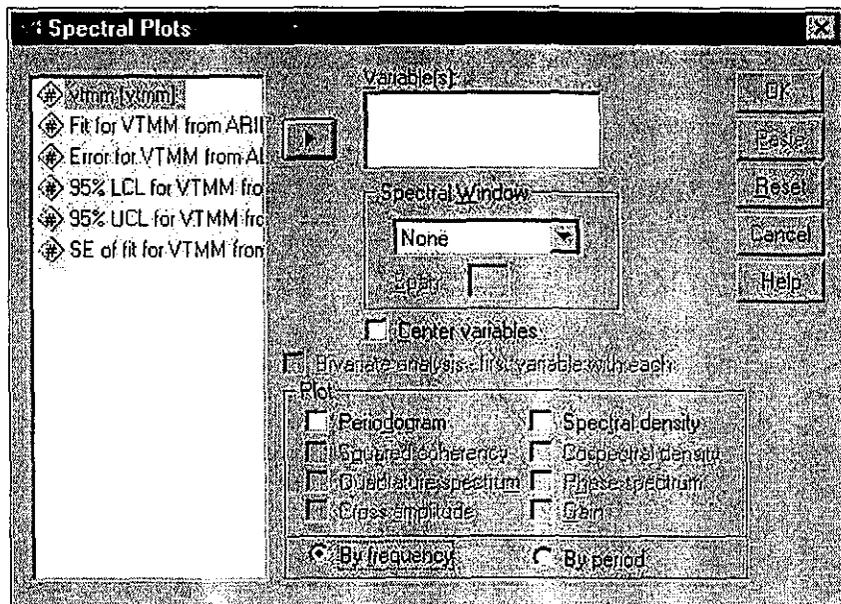


Figura A.XI. Cálculo de periodograma con SPSS.

Ejemplo

En la figura A.12 se muestra como se calculó el periodograma de los residuos Z_t , y que fue incluido como figura 4.21.

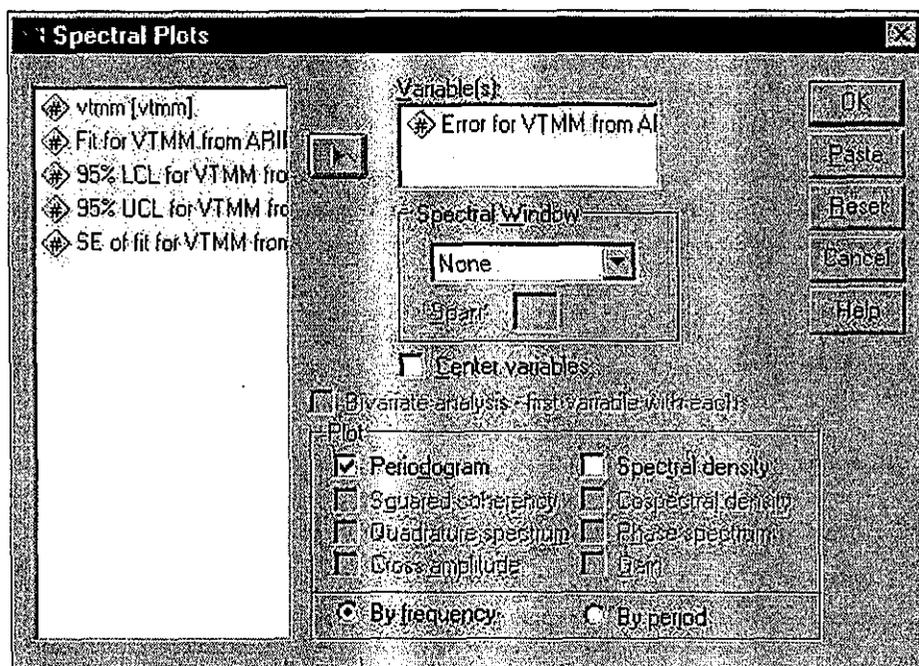


Figura A.XII. Cálculo del periodograma de los residuos Z_t .

TESIS CON
FALLA DE ORIGEN

BIBLIOGRAFÍA

1. **Chatfield**, Christopher; *The Analysis of Time Series: An Introduction*; 4ª. Ed.; Chapman and Hall; Gran Bretaña; 1991.
2. **Estrada**, Mónica; *Aplicación de Modelos Arima al Ozono en la Zona Suroeste de la Ciudad de México*; Tesina de Especialidad; UNAM; 1998.
3. **ALADI**, Asociación Latinoamericana de Integración; *Programa Regional de Estadísticas de Turismo*; OEA-ALADI; Montevideo; 1999.
4. **OMT**, Organización Mundial de Turismo; *Recomendaciones sobre Estadísticas de Turismo*; Informes Estadísticos, Serie M, No. 83; Naciones Unidas, Nueva York; 1994.
5. **SECTUR**, Secretaría de Turismo; *Compendio Estadístico del Turismo en México 2001*; www.mexico-travel.com.mx
6. **Box**, George E. P. & **Jenkins**, Gwilym M.; *Time Series Analysis, Forecasting and Control*; Holden Day; San Francisco; 1970.
7. **Chiang**, Chin L.; *An Introduction to Stochastic Processes and their Applications*; Robert E. Krieger; New York; 1968.
8. **Guerrero**, V.M.; *Análisis Estadístico de Series de Tiempo Económicas*; UAM; 1991.
9. **Brockwell**, Peter J. & **Davis**, Richard A.; *Introduction to Time Series and Forecasting*; Springer-Verlag; New York; 1966.