

Modelos Box-Jenkins

Aplicación de su metodología a la producción de azúcar en Cuba

MsC. Concepción Rodríguez Rodríguez *

El presente trabajo plantea la metodología de la aplicación de los modelos Box-Jenkins en datos reales de la producción de azúcar en Cuba en el periodo 1900-1998, donde se demuestra la presencia de cambios estructurales en la serie, mediante la prueba de Chow y se ajustan los modelos específicos para cada una de las etapas en la cual queda dividida la serie. Se verifican los supuestos en la etapa donde se observan las pruebas estadísticas aplicadas y las hipótesis verificadas. Se presenta además el esquema con los pasos a seguir en todo el proceso de identificación, estimación y verificación del modelo seleccionado.

LA EXPERIENCIA ha demostrado que la mayoría de los fenómenos reales de carácter, ya sea económico o de otra índole, son complejos y es necesario para una representación más real de esta, la aplicación de técnicas específicas que permitan una mayor flexibilidad a los modelos, introduciendo para estos una componente aleatoria en los modelos.

En el presente trabajo pretendemos proponer el desarrollo de la metodología de aplicación de los modelos Box-Jenkins, tomando como base los datos de una serie de producción de azúcar de 1900-1998, considerando las etapas que re-

* Profesora asistente de la Universidad de Granada.

quiere su tratamiento, así como el análisis de posibles cambios estructurales en la serie. Para el procesamiento de la información se utilizó el *software EViews*.

Presentamos además de algunos aspectos teóricos, los resultados obtenidos de la aplicación de esta técnica a una de las etapas obtenidas después del análisis del cambio estructural. El campo teórico relacionado con la aplicación de este modelo es muy amplio y complejo, pero la aplicación práctica de esta metodología constituye una herramienta muy eficaz para las investigaciones estadísticas y la realización de pronósticos.

¿En qué consiste la metodología Box-Jenkins?

Existen fenómenos reales que dada su complejidad no pueden ser representables mediante las ecuaciones en diferencias lineales, que aunque son de utilidad en la práctica imponen limitantes en lo que a su representación se refiere, debido a su característica de ser completamente deterministas, es por ello que resulta conveniente introducir en ellas una componente aleatoria que les permita una mayor flexibilidad.

De esta forma surgen los modelos:

- Autorregresivos [AR (p)].
- Medios Móviles [MA (q)].
- Mixtos [ARMA (p, q)].
- Autorregresivos integrados de medias móviles [ARMA (p, d, q)].

Estos últimos son modelos más generales producto de las combinaciones de los MA y AR. Este enfoque fue propuesto por Box y Jenkins (1970) y la idea fundamental radica en la estrategia que ellos proponen para *construir modelos*, los cuales no solo deben ser adecuados para representar el comportamiento de los datos observados, sino que la elección debe ser sugerida para los datos mismos, lo que se contrapone en el enfoque tradicional, que simplemente busca lograr el mejor ajuste de modelos preconcebidos según los datos que se manejen.

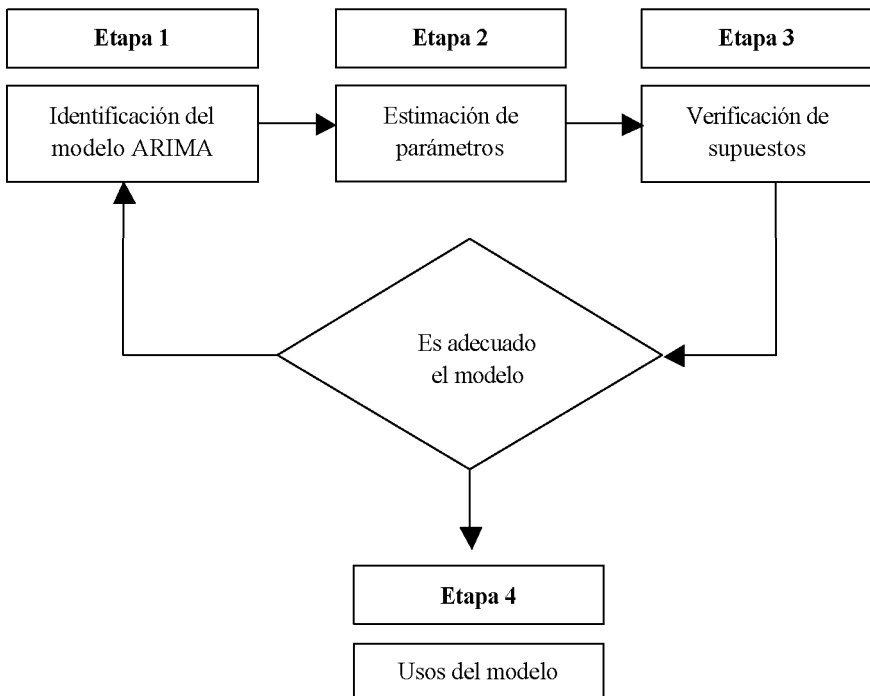
Los pasos a seguir para desarrollar esta metodología se exponen en el Anexo 1.

¿Qué pasos se sigue para la construcción del modelo?

La estrategia de construcción de modelos para series de tiempo desarrollado por Box y Jenkins (1970) consta de cuatro etapas fundamentales:

1. Identificación.
2. Estimación.
3. Verificación.
4. Uso del modelo.

Esta estrategia es un proceso iterativo que se representa en el siguiente esquema:



El objetivo principal es determinar primero una serie estacionaria en función de la serie original, para la cual se puede tener una representación ARMA (p , q) y posteriormente fijar los valores de p y q . Deberá analizarse el número de veces que se aplicará el operador diferencia después de haber determinado si es necesaria alguna transformación para estabilizar la varianza.

Un aspecto importante en la modelación ARIMA de una serie de tiempo simple es el número de veces que esta necesita de una diferencia antes de fijar el modelo. Para esto se utilizó la Prueba de Dickey Fuller o de raíces unitarias en la cual la hipótesis a verificar es:

Ho: Hay raíz unitaria (proceso no estacionario).

H1: No hay raíz unitaria (proceso estacionario).

El estadístico de prueba es:

$$T_u = \hat{a} / \hat{s}(\hat{a}_0)$$

Donde: \hat{a}_0 Estimación mínimo cuadrática de a_0

La regla de decisión a considerar será:

Rechazar H_0 si $|t_u| / t_{\alpha}$ y por tanto el proceso es estacionario.

Es estadístico t_u sigue una distribución t de Students que se tabuló por Dickey y Fuller. Los pasos para esta etapa de identificación se reflejan en el anexo 1.

Para la determinación del modelo es necesaria la observación visual del correlograma (f.a.c.m. y f.a.c.p.) de la variable diferenciada (en caso de que hubiera diferencia y en función del comportamiento de los picos se sugieren los modelos y se determinan los valores p , d y q).

De forma general para determinar si el modelo es el adecuado se utiliza la prueba de Box-Pierce que nos permite determinar en los modelos Box-Jenkins si el modelo determinado está correctamente especificado.

1. Estimación de parámetros

Una vez identificado el modelo adecuado se pasa a encontrar los mejores valores de los parámetros, a través de técnicas de estimación no lineal, para que dicho modelo represente apropiadamente a la serie considerada. Un método muy utilizado es el de Gaus-Newton. Para su

aplicación se necesita obtener estimaciones previas de los parámetros, tomadas como punto de partida para iniciar un proceso iterativo. En cada iteración se minimiza una aproximación lineal de la función procedente de un desarrollo de serie de Taylor en torno a la estimación correspondiente a la iteración anterior.

2. Verificación del modelo

Antes de utilizar el modelo seleccionado a fines de predicción debe someterse a prueba y cumplir algunos supuestos relativos a:

- a) los coeficientes;
- b) el término de error.

La etapa de verificación tiene su origen en la idea de que *todo modelo es erróneo*, puesto que son representaciones simples de la realidad. Una de las formas para detectar violaciones a los supuestos es a través del análisis de los residuales $\{a_t\}$, que representan la parte de las observaciones que no es explicada por el modelo.

Se plantean ocho supuestos para la verificación del modelo donde los cinco primeros corresponden al término de error y los tres restantes a los coeficientes.

Supuesto 1. –La serie de los a_t tiene media cero.

Supuesto 2. –La serie de los a_t tiene varianza constante.

Supuesto 3. –Las variaciones aleatorias son mutuamente independientes.

Supuesto 4. –Los a_t se distribuyen normalmente para toda t .

Supuesto 5. –No existen observaciones aberrantes.

Supuesto 6. –El modelo es parsimonioso o cumple el principio de parquedad.

Supuesto 7. –El modelo es admisible.

Supuesto 8. –El modelo es estable en los parámetros.

3. Usos del Modelo.

Está en función de los fines que persigue el investigador con su construcción, los cuales para lo general son pronósticos, control, simulación o explicación del fenómeno en estudio.

Consideraciones acerca de la serie analizada

La serie utilizada abarca el período 1900-1998 donde se observa que hay un número considerable de datos y el indicador estudiado se afecta sensiblemente por los cambios históricos-político y sociales del país, lo que conlleva que al graficar dicha serie se observe la presencia de diferentes comportamientos y podemos suponer la existencia de cambios estructurales, y se verifica aplicando la prueba de Chow a dicha serie, obteniendo los resultados siguientes:

Sample break point(s): 14

F-Statistic	5,50516	Probability	0,0056
Likelihood ratio	10,8472	Probability	0,0044

Sample break point(s): 34

F-Statistic	13,9478	Probability	0,0000
Likelihood ratio	25,3559	Probability	0,0000

Sample break point(s): 63

F-Statistic	3,60485	Probability	0,0312
Likelihood ratio	7,24413	Probability	0,0267

Sample break point(s): 90

F-Statistic	4,70958	Probability	0,0114
Likelihood ratio	9,35572	Probability	0,0093

En los puntos de ruptura planteados se observa que en todos los casos la $P < 0,05$ lo que prueba que se rechaza H_0 : No hay cambio estructural y se puede concluir que la serie se subdivide en cinco etapas:

1. 1900-1912.
2. 1913-1932.
3. 1933-1962.
4. 1963-1989.
5. 1990-1998.

Por lo extenso que resultaría el análisis de todas las etapas analizaremos solo la etapa 1963-1989 para ilustrar la aplicación del método.

En esta etapa se determinó no considerar el año 1970 por ser un año atípico en toda la serie y no refleja el real comportamiento de este indicador.

Análisis del período 1963-1989

Según el gráfico de la serie se observa una tendencia creciente con fluctuaciones frecuentes producto de la presencia de variaciones estructurales y organizativas al menos hasta 1971, percibiéndose luego una recuperación económica discreta con un brusco descenso en 1979.

Para comenzar la etapa de identificación se corre el correlograma para la serie analizada donde se observa que la tendencia de la f.a.s. a cero es bastante lenta por lo que se considera que el proceso es no estacionario.

Con el objetivo de eliminar la tendencia y hacer la serie estacionaria se aplicó una diferencia a la serie original y se obtiene la variable DPROD3 la cual mediante la prueba de raíces unitarias de Dickey-Fuller se verificó que es el orden adecuado de la diferencia según los resultados siguientes:

Augmented Dickey-Fuller: UROOT (T,1) DPROD3

<i>Dickey-Fuller:</i>	<i>t-Statistic</i>	<i>-6,0308</i>
<i>Mackinnon critical values</i>	<i>1 %</i>	<i>-4,4167</i>
	<i>5 %</i>	<i>-3,6219</i>
	<i>10 %</i>	<i>-3,2474</i>

Augmented Dickey-Fuller: UROOT (T,1) DPROD3

<i>Dickey-Fuller:</i>	<i>t-Statistic</i>	<i>-6,0788</i>
<i>Mackinnon critical values</i>	<i>1 %</i>	<i>-2,6700</i>
	<i>5 %</i>	<i>-1,9566</i>
	<i>10 %</i>	<i>-1,6235</i>

Al graficar la variable diferenciada se observó estabilidad en la varianza, así como al obtener los valores de la S de la serie original (S_0) y diferenciada (S_1) se observó una disminución de esta.

Para la determinación del modelo se calculan las f.a.s y f.a.c.p donde según se observa en el correlograma para la variable diferenciada podemos tomar las siguientes consideraciones:

- Existe un pico predominante en la primera autocorrelación muestral y parcial, que puede sugerir un ARIMA (1,1,1).
- En la f.a.c.p se observa también un pico acentuado en la sexta autocorrelación, que sugiere un ARIMA (1,1,6).

En base a estos elementos se corrió la regresión para estimar los valores de los parámetros del modelo que mejor represente el comportamiento de la serie:

Se probó la regresión para los siguientes modelos:

- a) ARIMA (1,1,1): los coeficientes no resultan significativos.
- b) ARIMA (1,1,6): el componente AR no da significativo, por lo que se prueba un MA.
- c) IMA (1,1): Observándose que: el coeficiente es significativo, pero no cumple la condición de invertibilidad.
- d) ARI (1,1): el coeficiente es significativo y cumple la condición de estacionalidad.

El estadístico F es también significativo decidiéndose adoptar el modelo ARI (1,1).

Las salidas de la regresión se muestran a continuación:

LS // Dependent variable is DPROD3

Variable	Coefficiente	STD-Error	T-Stat	2-TAIL SIG
C	129,73133	16893,981	0,0076791	0,9939
MA (1)	-1,0975466	0,0026895	-408,08784	0,0000
AR (1)	0,5491310	3,4905080	0,1573212	0,8765

LS // Dependent variable is DPROD3

Variable	Coefficiente	STD-Error	T-Stat	2-TAIL SIG
C	164,64065	195,05662	0,8440659	0,4111
MA (1)	-0,9339117	0,2403914	-3,8849629	0,0013
AR (6)	-0,3414439	0,1468303	-2,3234329	0,0335

LS // Dependent variable is DPROD3

Variable	Coefficiente	STD-Error	T-Stat	2-TAIL SIG
C	151,33900	384,25329	0,3938522	0,6973
MA (1)	-1,023900	0,0993780	-10,303875	0,0000

LS // Dependent variable is DPROD3

Variable	Coefficiente	STD-Error	T-Stat	2-TAIL SIG
C	150,27089	129,04730	1,1644636	0,2567
AR (1)	-0,5434498	0,1794561	-3,0283159	0,0062

F-Statistic 9,170697 Prob (F-Statistic) 0,006175

El modelo propuesto será: $Y_t = 0,54344 Y_{t-1} + a_t$

Este modelo pasa a la etapa de verificación donde se analizan los coeficientes y los residuales mediante los supuestos planteados.

Supuesto 1

Ho: $\mu (a_t) = 0$

H₁: $\mu (a_t) \neq 0$

Si $m(\hat{a}) / s(\hat{a}) \sqrt{n - d - p - q} \leq 2$ Acepto H₀

Los valores de $m(\hat{a}) = 17,669181$ y $s(\hat{a}) = 938,39365$ se toman de la salida del histograma de los residuos. Se calcula el coeficiente y obtenemos que

$/0,00503/ < 2$, por lo que aceptamos H_0 y se concluye que la media de los residuales es cero.

Supuesto 2

Se utilizó la prueba de ARCH (heterocedasticidad condicionada a los errores autorregresivos) para verificar este supuesto, en las hipótesis a plantear son:

H_0 : No hay efectos ARCH

H_1 : Hay efectos ARCH

ARCH	Test
Obs * Squared:	4,20537
Probability:	0,7558

Los resultados mostrados indican que el estadístico NR^2 tiene una probabilidad marginal de 0,7558 por lo cual se acepta H_0 y se concluye que al no existir heterocedasticidad condicionada a los errores autorregresivos podemos asegurar que los residuales tienen varianza constante.

Supuesto 3

El estadístico de Ljung-Box obtenido en el correlograma de residuos nos permite verificar el supuesto de independencia para toda K diferente de cero, el cual toma el valor de 9,44 con una $P = 0,2228$, por lo que podemos concluir que los residuales son estadísticamente independientes.

Supuesto 4

En el histograma realizado se observa la simetría de la serie obteniéndose los siguientes resultados:

Jarque-Bera normality test stat	0,7280	Probability:	0,644891
---------------------------------	--------	--------------	----------

Las hipótesis a verificar son:

H_0 : los residuales se distribuyen normalmente;

H_1 : los residuales no se distribuyen normalmente.

Por lo que podemos decir que se acepta H_0 y se concluye que los residuales se distribuyen normalmente.

Supuesto 5

Al graficar los residuales se observa que todos los valores obtenidos caen dentro del intervalo $[\pm 3s(\hat{a})]$ que toma valores $[-2815,181; 2815,181]$ y solo un punto cae fuera del intervalo $[\pm 2s(\hat{a})]$ por lo que se concluye que no hay observación aberrante.

Supuesto 6

Al calcular el intervalo de confianza para el parámetro $[f \pm 2s(f)]$ para un 95 % de confiabilidad se obtiene $[-0,90236; 0,18400]$ observándose que dicho intervalo no contiene el cero, lo que indica que el parámetro explica el comportamiento del indicador y el modelo cumple el principio de parquedad.

Supuesto 7

La condición de estacionalidad de un AR (1) es $|f| < 1$, la cual se cumple ya que $|f_1| = 0,54344$, lo que indica que el modelo es admisible.

Supuesto 8

Este supuesto no se analiza ya que el modelo es un AR (1,1) y solo contiene un parámetro. De esta forma queda verificada que el modelo propuesto $Y_t = -0,54344 Y_{t-1} + a_t$ cumple con los requisitos y nos permite realizar pronósticos.

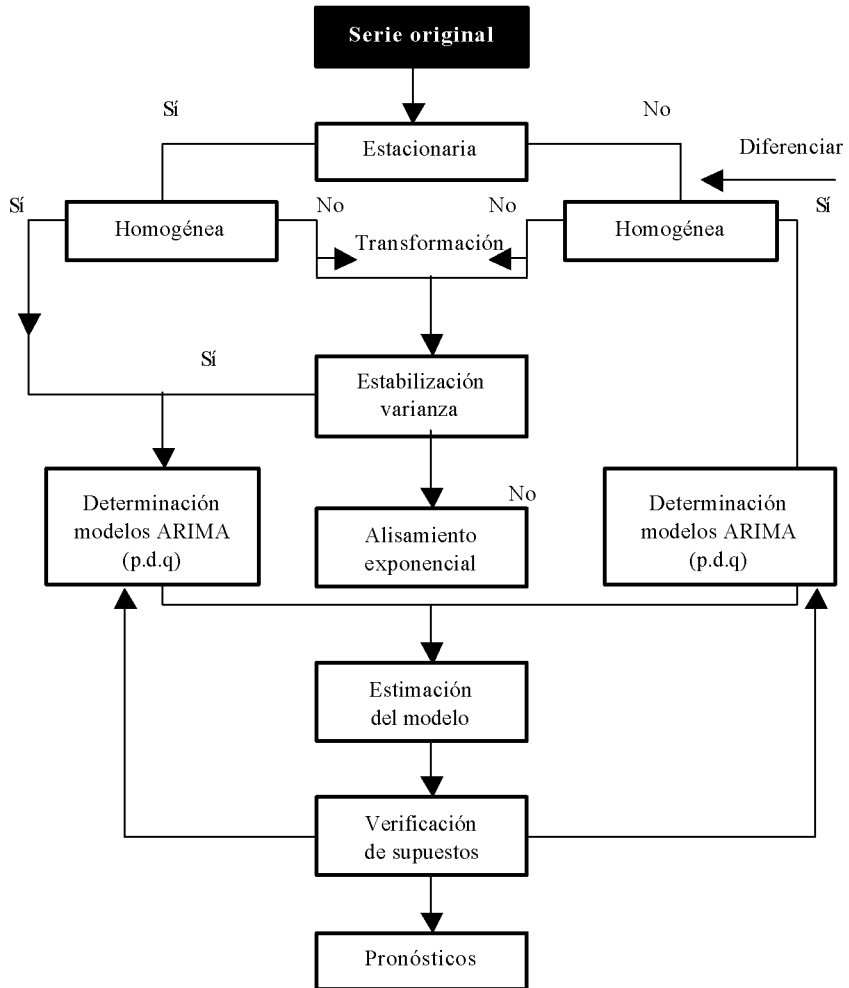
Al realizar el mismo análisis para el resto de las etapas se obtuvo el modelo adecuado para cada uno. Los resultados se muestran en la tabla del Anexo 2 así como la verificación de los supuestos.

Bibliografía

- Damodar No. Gujarate: *Econometría*. Segunda edición. Mc. Graw Hill, Colombia, 1990.
- Guerrero Víctor, M: *Análisis Estadístico de Series de tiempo Económicas*. Colección BI-México, 1991.
- Harvey Andrews: *The econometrics analysis on time series*. Segunda edición, The Mit Press, Cambridge, Massachusetts, 1990.
- Johnston I: *Econometrics Methods*. Mc. Graw Hill, Thind edition, 1984.
- Kennedy Peter: *A guide to econometrics*. The Met Press, Cambridge, Massachusetts, 1992.
- Silva L. Arnaldo: *Cuba y el mercado internacional azucarero*. Editorial de Ciencias Sociales, La Habana, Cuba.
- Manual MICROTSP*. Setup Guide, Versión 7.0, Copyright, California, 1990.

Anexo 1

Esquema de pasos a seguir en la modelación Box-Jenkins



Anexo 2

Tabla No. 1

Cuadro resumen de los resultados obtenidos en el análisis de la serie por etapas

Período y número de observaciones	Modelos seleccionados	Parámetros seleccionados	Intervalos de confianza para el 95 %	Correlación entre parám. $>0,5$ ó $<0,5$	m (â) y cociente	P Arch n R^2 probab.	Q _{L-B} probab.	Estadist. J-B probab.	Oservaciones Atípicas
1913-1932 N = 20	ARIMA (2,1,2)	$F_2 = -0,9999$ $F_2 = 0,80655$	(-1,75211; -0,64797) (0,51486; 1,09824)	-0,23861	1,12659 0,007574	5,07505 0,53420	0,43000 0,99860	1,70817 0,42567	_____
1933-1962 N = 30	IMA (1,1)	$F_1 = -0,51670$	(-0,86678; -0,16662)	_____ -0,0003291	-0,05057 -0,0003291	5,82054 0,66730	3,78000 0,87670	0,00216 0,99892	_____
1963-1989 N = 26	ARI (1,1)	$F_1 = -0,54345$	(-0,90236; -0,18454)	_____ 0,09030156	17,66918 0,09030156	4,20537 0,75580	9,44000 0,22280	0,72800 0,69489	_____