

Identificación de un proceso ARIMA contaminado

Elkin Castaño V.

Introducción, 51. I. El proceso de identificación, 52. II. Aplicación del procedimiento a una serie simulada, 54. Conclusiones, 67. Apéndice, 68. Referencias, 70.

Introducción

Uno los problemas más frecuentes en la modelación de series de tiempo es la presencia de observaciones atípicas generadas por diferentes tipos de intervenciones. Como se sabe, dichas observaciones pueden distorsionar la identificación del proceso ARIMA generador de los datos realizada a través de las herramientas usuales como son la función de autocorrelación, la función de autocorrelación parcial, etc.

Este artículo propone un procedimiento de identificación relativamente simple el cual hace uso de los algoritmos de detección de observaciones atípicas Chang, Tiao y Chen -1988- o Chen y Liu -1990- implementados en el paquete estadístico SCA/UTS o SCA/XUTS.

La dificultad inicial de identificar un proceso, sobre el cual se aplica la detección de observaciones atípicas, es resuelta usando el hecho de que todo proceso ARIMA estacionario e invertible puede ser escrito como un proceso autorregresivo puro de alto orden o un proceso de media móviles puro de alto orden.

El plan de este documento es el siguiente: la sección I presenta el proceso de identificación propuesto, mientras que en la sección II el procedimiento es aplicado a una serie simulada; el apéndice contiene una macro, escrita para el paquete estadístico SCA, con los comandos básicos para el proceso de identificación aplicado a la serie simulada.

I. El proceso de identificación

Suponga que Z_t es una serie de tiempo con m intervenciones y que sigue el proceso:

$$Z_t = \sum_{j=1}^m v_j(B) I_t^{(T_j)} + N_t$$

donde

$$N_t = \frac{\theta(B)}{\delta(B)\phi(B)} a_t$$

donde $I_t^{(T_j)}$ es una variable dicótoma con un uno en el período T_j -ésimo, $j=1,2,\dots,m$, $v_j(B)$ es la función de transferencia correspondiente a la j -ésima intervención, B es el operador de rezagos usual, $\theta(B)$ es el polinomio de medias móviles con todas sus raíces fuera del círculo unidad, $\phi(B)$ es el polinomio autorregresivo con sus raíces fuera del círculo unidad y que no tiene factores comunes con $\theta(B)$ y $\delta(B)$ es el polinomio de diferencias con sus raíces sobre el círculo unidad.

Bajo estas condiciones el proceso de identificación es el siguiente:

i) Aproxime a Z_t usando un proceso puro autorregresivo o de medias móviles de orden alto. En la elección del orden se debe tener en cuenta la frecuencia del período de observación, y la clase de proceso -estacional o no. En otras palabras Z_t puede aproximarse como:

$$Z_t = \frac{\theta(B)}{\delta(B)} a_t$$

usando un proceso de medias móviles puro, o

$$Z_t = \frac{1}{\delta(B)\phi(B)} a_t$$

en términos de un proceso autorregresivo puro.

Estime Z_t y haga un análisis de residuales para verificar la buena aproximación a la estructura de Z_t del modelo estimado.

Si los residuales se comportan como ruido blanco el orden elegido aproxima adecuadamente la estructura del proceso.

ii) Detecte observaciones atípicas empleando el algoritmo de Chang y Chen -1988- o Chen y Liu -1990-. Estos algoritmos se encuentran implementados en el paquete estadístico SCA/UTS o SCA/XUTS. Si T es moderado o grande se recomienda usar un valor crítico de 3.5 o mayor. Este procedimiento arroja como resultado las diferentes clases de intervenciones, el período de ocurrencia y una estimación de su efecto.

iii) Suponga que en el período $T_j, j=1,2,\dots,m$ ocurrió la j -ésima intervención sobre la serie y que su estructura es la función de transferencia $v_j(B)$. Una estimación más refinada de los efectos de las intervenciones que las obtenidas en ii) puede obtenerse usando el modelo:

$$Z_t = \sum_{j=1}^m v_j(B) I_t^{(T_j)} + \frac{\theta(B)}{\delta(B)} a_t$$

o el modelo

$$Z_t = \sum_{j=1}^m v_j(B) I_t^{(T_j)} + \frac{1}{\delta(B)\phi(B)} a_t$$

iv) Sea $v_j^*(B)$ la estimación de $v_j(B)$ obtenida en iii). Una aproximación al proceso N_t puede conseguirse de la filtración de Z_t por la función

$$Z_t = \sum_{j=1}^m v_j^*(B) I_t^{(T_j)}$$

Es decir, una aproximación N_t^* al proceso N_t se logra descontaminando a Z_t de las intervenciones ocurridas. Por tanto N_t es aproximadamente:

$$N_t^* = Z_t - \sum_{j=1}^m v_j^*(B) I_t^{(T_j)}$$

La serie N_t^* es la filtración del proceso Z_t por las intervenciones ocurridas.

v) Sobre N_t^* utilice el procedimiento usual sugerido por Box y Jenkins para la identificación de procesos ARIMA.

Suponga que N_t^* es identificado como el proceso

$$N_t^* = \frac{\theta(B)}{\delta(B)\phi(B)} a_t$$

vi) Estime el modelo completo

$$Z_t = \sum_{j=1}^m v_j(B) I_t^{(T_j)} + \frac{\theta(B)}{\delta(B)\phi(B)} a_t$$

La próxima sección contiene la aplicación del procedimiento a una serie simulada.

II. Aplicación del procedimiento a una serie simulada

A continuación se aplicará el procedimiento una serie MA(1) simulada la cual ha sido intervenida en tres ocasiones usando diferentes clases de funciones de transferencia: una aditiva, una de cambio de nivel y una innovativa.

Suponga que Z_t es un proceso MA(1) con $T=150$ y que se encuentra intervenido en los siguientes períodos:

a) Período $T_1=50$ con una estructura de cambio de nivel, es decir,

$$v_1(B) = WLS / (1 - B),$$

donde WLS es un parámetro que mide el impacto inicial de la intervención; otra forma de escribir esta función de transferencia es $v_1 = WLS$, pero donde la variable $I_t^{(T_1)}$ es ya una variable indicadora con ceros antes del período T_1 y unos desde dicho período.

b) Período $T_2 = 90$ con una estructura innovativa, es decir,

$$v_2(B) = WIO(1 - \theta B),$$

donde WIO es un parámetro que mide el impacto inicial de la intervención.

c) Período $T_3 = 125$ con una estructura aditiva, es decir,

$$v_3(B) = WAO,$$

donde WAO mide el impacto inicial de la intervención.

Con esto, el proceso generador de los datos es de la forma:

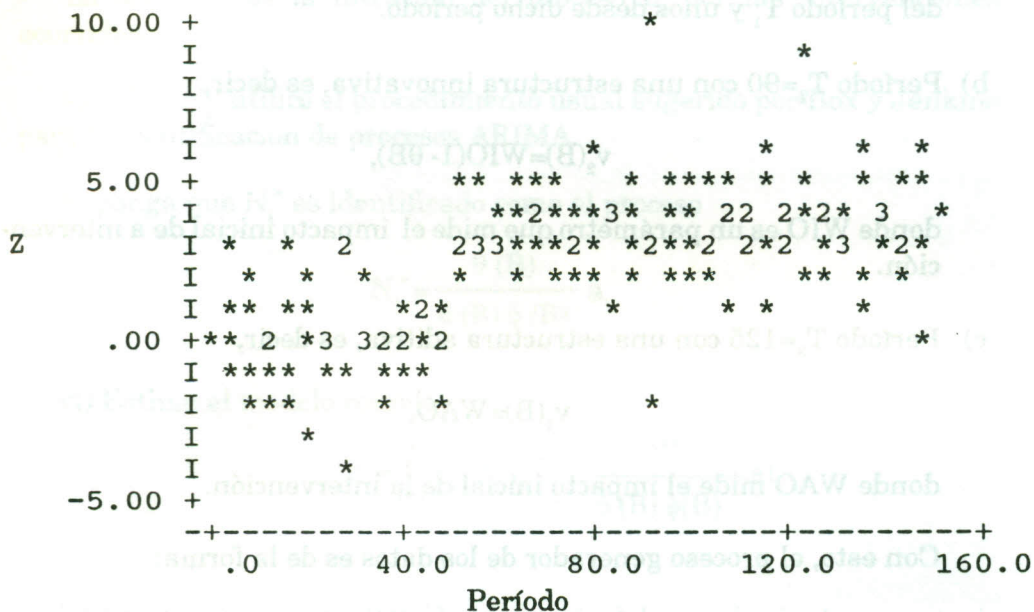
$$Z_t = \sum_{j=1}^3 v_j(B) I_t^{(T_j)} + (1 - \theta B) a_t$$

donde las funciones de transferencia v_j ya fueron especificadas, $I_t^{(T_j)}$, $j=2,3$, son variables indicadoras con un uno en los períodos 90 y 125, $I_t^{(T_1)}$ es una variable indicadora con ceros antes del período 50 y unos desde dicho período y a_t es ruido blanco con distribución normal de media cero y varianza uno.

Los valores para los parámetros del modelo son los siguientes:

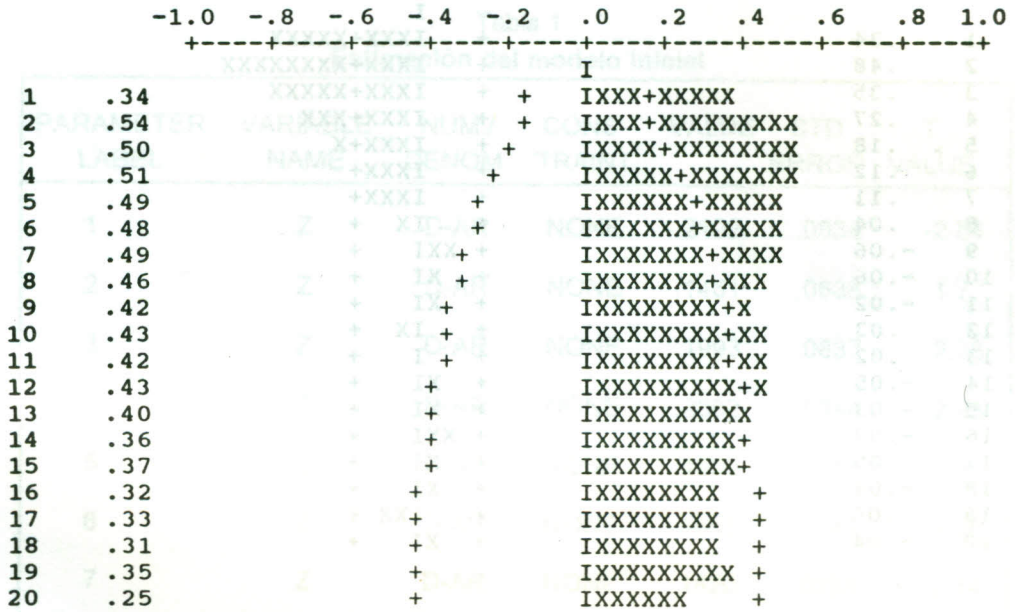
WAO=4.8, WLS=3.5, WIO=7.5 y $\theta = 0.85$. La serie simulada se presenta en la gráfica 1.

Gráfica 1
Serie Simulada



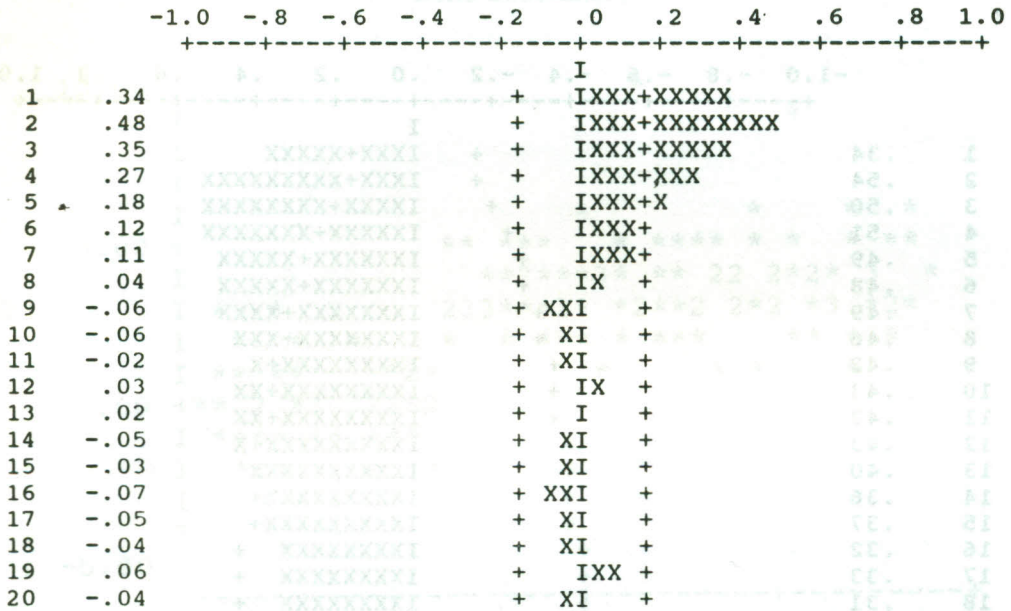
Si no hacemos caso a las intervenciones que han ocurrido sobre Z_t y empleamos el procedimiento de identificación usual sugerido por Box y Jenkins se obtienen funciones de autocorrelación y autocorrelación parcial muestrales dadas en las gráficas 2 y 3.

Gráfica 2
Autocorrelaciones de la serie intervenida



Como se aprecia en estos resultados, la información de que el proceso ARIMA es un MA1 se ha perdido en las funciones de autocorrelación debido a los efectos de las intervenciones. La interpretación directa de ellos conduciría muy posiblemente a identificar un modelo espúreo para Z_t .

Gráfica 3
Autocorrelaciones parciales de la serie intervenida



Identificación de la serie contaminada usando el procedimiento propuesto

A continuación aplicaremos el procedimiento de identificación propuesto en la sección I.

i) Definición del modelo inicial

Como se dijo en la sección I, para la definición del modelo inicial deben tenerse en cuenta las características del proceso, como por ejemplo la estacionalidad y frecuencia de la serie. La buena aproximación del modelo inicial puede ser verificada por medio del análisis de residuales de dicho modelo ajustado.

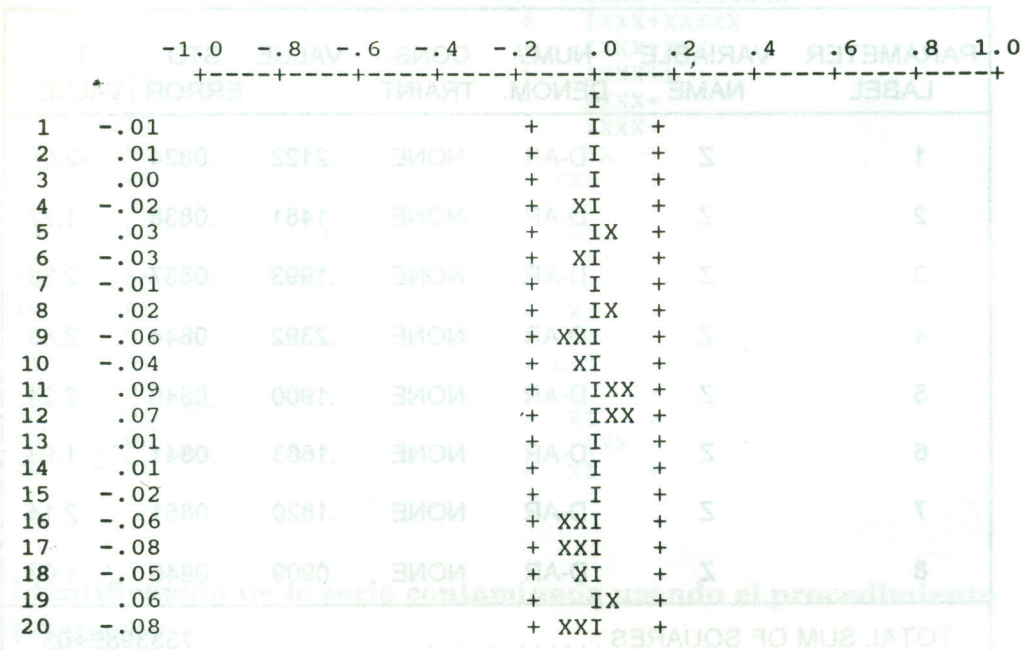
En nuestro caso, supondremos como modelo inicial un modelo autorregresivo de orden 8, es decir un AR(8). La estimación arrojó los siguientes resultados.

Tabla 1
Estimación del modelo inicial

PARAMETER LABEL	VARIABLE NAME	NUM./ DENOM.	CONS- TRAI NT	VALUE	STD ERROR	T VALUE
1	Z	D-AR	NONE	-.2122	.0834	-2.54
2	Z	D-AR	NONE	.1481	.0838	1.77
3	Z	D-AR	NONE	.1993	.0837	2.38
4	Z	D-AR	NONE	.2392	.0844	2.83
5	Z	D-AR	NONE	.1900	.0846	2.24
6	Z	D-AR	NONE	.1663	.0841	1.98
7	Z	D-AR	NONE	.1820	.0851	2.14
8	Z	D-AR	NONE	.0909	.0846	1.07
TOTAL SUM OF SQUARES753398E+03	
TOTAL NUMBER OF OBSERVATIONS						150
RESIDUAL SUM OF SQUARES351914E+03	
R-SQUARE507
EFFECTIVE NUMBER OF OBSERVATIONS						142
RESIDUAL VARIANCE ESTIMATE247827E+01	
RESIDUAL STANDARD ERROR157425E+01	

La gráfica 4 muestra el correlograma muestral de los residuales del modelo ajustado.

Gráfica 4
Autocorrelaciones del modelo inicial ajustado



De estos resultados parece ser que la especificación del modelo inicial es adecuada.

ii) Detección de observaciones atípicas.

La aplicación del algoritmo de Chang, Tiao y Chen -1988- utilizando un valor crítico de 3.5 arroja los siguientes resultados

INITIAL RESIDUAL STANDARD ERROR = 1.5691

TIME	ESTIMATE	T-VALUE	TYPE
50	3.43	4.94	LS
90	6.12	4.55	IO
125	5.43	4.83	AO

ADJUSTED RESIDUAL STANDARD ERROR = 1.1476

El procedimiento ha identificado correctamente las intervenciones simuladas sobre Z_t . En efecto, en el período 50 identificó una intervención de cambio de nivel (LS); en el período 90 encontró una intervención innovativa (IO) y, finalmente, en el período 125 identificó una intervención aditiva (AO).

iii) Estimación de los efectos de las intervenciones.

Para conseguir estimadores más eficientes de los efectos de las intervenciones se ajusta el modelo intervenido AR(8):

$$Z_t = wlsI_t^{(T_1)} + \frac{wio}{1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_8 B^8} I_t^{(T_2)} + waoI_t^{(T_3)} + \frac{1}{1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_8 B^8} a_t$$

donde la variable $I_t^{(T_1)}$ contiene ceros antes del período 50 y unos a partir de dicho período, $I_t^{(T_2)}$ contiene un uno en el período 90 y ceros en los demás e $I_t^{(T_3)}$ contiene un uno en el período 125 y cero en los demás.

NOTA: La especificación del modelo en el paquete SCA es la siguiente:

TSMODEL MOD1. MODEL Z= (WAO)INT125AO(BINARY)+ @
 (WLS)INT50LS(BINARY)+ @
 (WIO)/(1-P1*B-P2*B**2-P3*B**3- @
 P4*B**4-P5*B**5-P6*B**6- @
 P7*B**7-P8*B**8)INT90IO(BINARY)+ @
 1/(1-P1*B-P2*B**2-P3*B**3-P4*B**4- @
 P5*B**5-P6*B**6-P7*B**7-P8*B**8)NOISE. @

dónde los $P_j, j=1, \dots, 8$, son los parámetros autorregresivos.

Los resultados de las estimación se presentan en la Tabla 2.

La gráfica 5 presenta el correlograma para los residuos del modelo.

Gráfica 5
 Autocorrelaciones del modelo inicial interv.

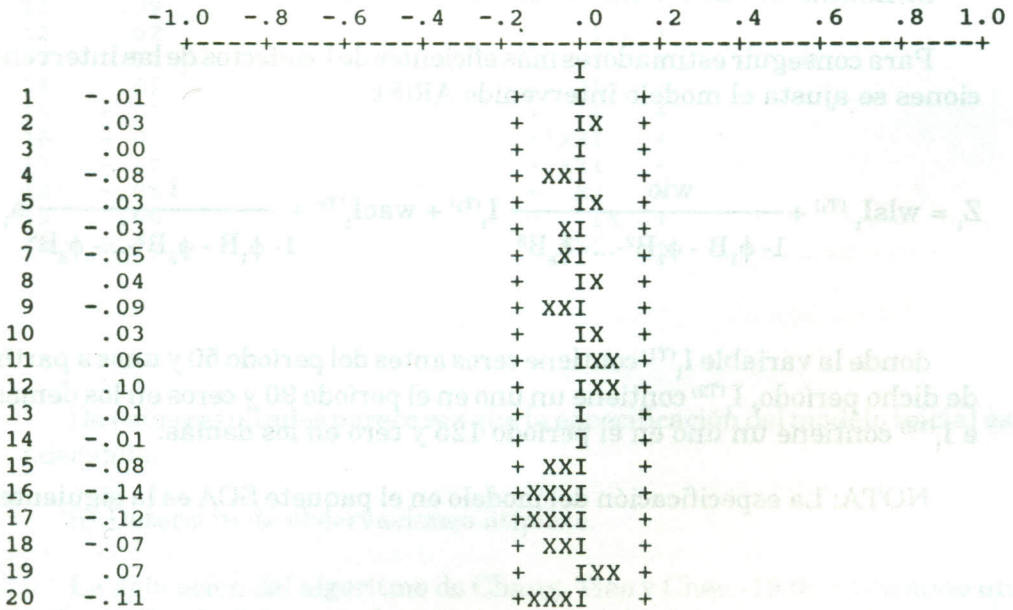


Tabla 2
Estimación del modelo inicial con intervenciones

PARAMETER LABEL	VARIABLE NAME	NUM./ DENOM.	CONS- TRAINT	VALUE	STD ERROR	T VALUE
1	WAO	INT125AO	NUM. NONE	5.0665	.6274	8.08
2	WLS	INT50LS	NUM. NONE	3.4747	.0259	133.99
3	WIO	INT90IO	NUM. NONE	6.5793	1.0038	6.55
4	P1	INT90IO	DENM EQ 01	-.7984	.0733	-10.89
5	P2	INT90IO	DENM EQ 02	-.6054	.0936	-6.47
6	P3	INT90IO	DENM EQ 03	-.5392	.1007	-5.36
7	P4	INT90IO	DENM EQ 04	-.3745	.1030	-3.64
8	P5	INT90IO	DENM EQ 05	-.3251	.1031	-3.15
9	P6	INT90IO	DENM EQ 06	-.2018	.1017	-1.98
10	P7	INT90IO	DENM EQ 07	-.0882	.0936	-.94
11	P8	INT90IO	DENM EQ 08	-.0401	.0739	-.54
***	P1	Z	D-AR EQ 01	-.7984	.0733	-10.89
***	P2	Z	D-AR EQ 02	-.6054	.0936	-6.47
***	P3	Z	D-AR EQ 03	-.5392	.1007	-5.36
***	P4	Z	D-AR EQ 04	-.3745	.1030	-3.64
***	P5	Z	D-AR EQ 05	-.3251	.1031	-3.15
***	P6	Z	D-AR EQ 06	-.2018	.1017	-1.98
***	P7	Z	D-AR EQ 07	-.0882	.0936	-.94
***	P8	Z	D-AR EQ 08	-.0401	.0739	-.54
TOTAL SUM OF SQUARES753398E+03	
TOTAL NUMBER OF OBSERVATIONS						150
RESIDUAL SUM OF SQUARES134998E+03	
R-SQUARE811
EFFECTIVE NUMBER OF OBSERVATIONS						142
RESIDUAL VARIANCE ESTIMATE950690E+00	
RESIDUAL STANDARD ERROR975034E+00	

iv) Filtración de la serie contaminada por las intervenciones detectadas

A continuación de filtra la serie contaminada Z_t usando el modelo de las intervenciones estimadas en el paso anterior.

En el paquete SCA se haría de la siguiente forma -wls, wio, wao y P1 a P8 son los valores estimados obtenidos en la etapa anterior:-

```
TSMODEL MOD2. MODEL Z= (WAO)INT125AO(BINARY)+ @
(WLS)INT50LS(BINARY)+ @
(WIO)/(1-P1*B-P2*B**2-P3*B**3- @
P4*B**4-P5*B**5-P6*B**6- @
P7*B**7-P8*B**8)INT90IO(BINARY)+NOISE.
```

```
FILTER OLD Z. NEW ZF. MODEL MOD2.
```

v) Identificación del modelo ARIMA original

Sobre la serie filtrada empleamos el procedimiento tradicional de identificación de Box y Jenkins. Esto arroja los resultados presentados en las gráficas 6 y 7.

El procedimiento de indentificación muestra que el probable proceso generador de los datos es un MA(1), lo cual coincide con el modelo especificado en la simulación.

vi) Estimación del modelo final

El modelo final a estimar será de la forma:

$$Z_t = wlsI_t^{(T_1)} + wio (1-\theta B) I_t^{(T_2)} + waoI_t^{(T_3)} + (1-\theta B)a_t$$

La especificación de este modelo en el paquete estadístico SCA es la siguiente

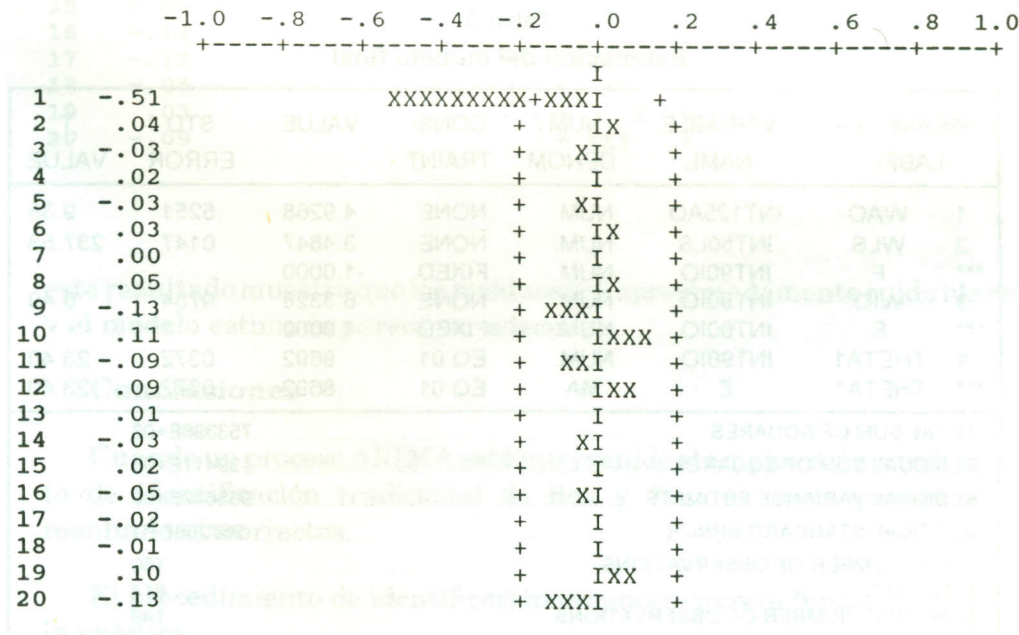
F=-1

TSMODEL MODMA4. MODEL Z= (NWA0)INT125AO(BINARY)+ @
 (NWLS)INT50LS(BINARY)+ @
 (F)(NWIO)(F+THETA1*B)INT90IO(BINARY)+ @
 (1-THETA1*B)NOISE. FIXED F.

Observe el manejo que se debe dar a la parte de la intervención innovativa. Esto se debe a que SCA trata en forma diferente el numerador de la función de transferencia y el numerador -polinomio de medias móviles- del modelo del término de error.

Los resultados de la estimación se encuentran en la tabla 3.

Gráfica 6
 Autocorrelaciones de la serie filtrada



Gráfica 7
Autocorrelaciones parciales de la serie filtrada

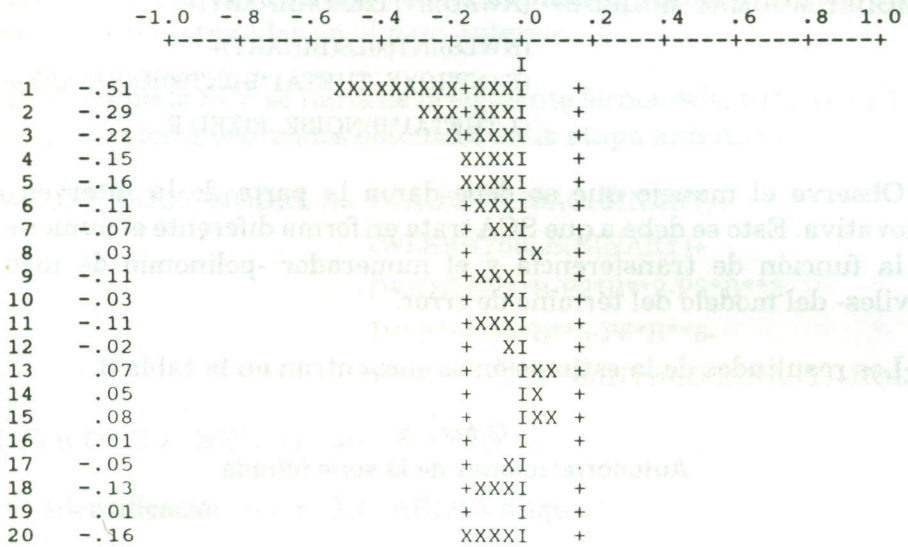
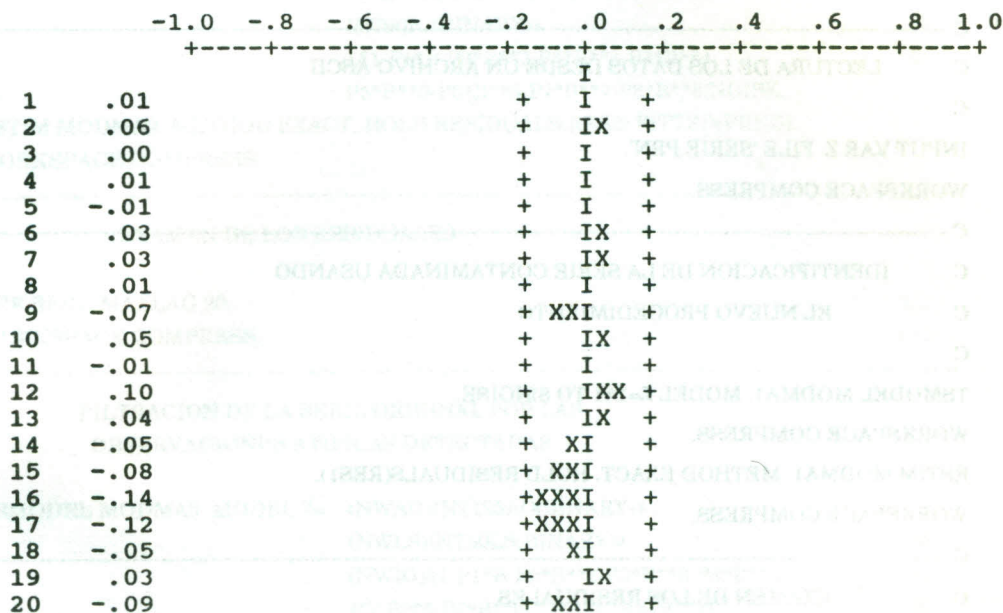


Tabla 3
Estimación del modelo final

PARAMETER LABEL	VARIABLE NAME	NUM./ DENOM.	CONS- TRAI NT	VALUE	STD ERROR	T VALUE	
1	WAO	INT125AO	NUM.	NONE	4.9268	.5251	9.38
2	WLS	INT50LS	NUM.	NONE	3.4847	.0147	237.53
***	F	INT90IO	NUM.	FIXED	-1.0000		
3	WIO	INT90IO	NUM.	NONE	6.3328	.9754	6.49
***	F	INT90IO	NUM.	FIXED	-1.0000		
4	THETA1	INT90IO	NUM.	EQ 01	.8692	.0372	23.40
***	THETA1	Z	MA	EQ 01	.8692	.0372	23.40
TOTAL SUM OF SQUARES753398E+03		
RESIDUAL SUM OF SQUARES					139411E+03		
RESIDUAL VARIANCE ESTIMATE					935647E+00		
RESIDUAL STANDARD ERROR967288E+00		
TOTAL NUMBER OF OBSERVATIONS					150		
R-SQUARE814		
EFFECTIVE NUMBER OF OBSERVATIONS					149		

La gráfica 8 presenta el correlograma de los residuales del modelo final.

Gráfica 8
Autocorrelaciones del modelo final



este resultado muestra que los residuos son aproximadamente ruido blanco y el modelo estimado parece ser adecuado.

Conclusiones

Cuando un proceso ARIMA está intervenido el empleo del procedimiento de identificación tradicional de Box y Jenkins suele proporcionar resultados incorrectos.

El procedimiento de identificación propuesto parece funcionar bien en la práctica.

Apéndice

Marco para la identificación del proceso contaminado del ejemplo

==SIM

PROFILE IWIDTH 80. REVIEW.

C -----

C LECTURA DE LOS DATOS DESDE UN ARCHIVO ASCII

C 10. 1

INPUT VAR Z. FILE 'SERIE.PRN'. 00. 2

WORKSPACE COMPRESS. 01. 3

C -----

C IDENTIFICACION DE LA SERIE CONTAMINADA USANDO

C EL NUEVO PROCEDIMIENTO

C 01. 4

TSMODEL MODMA1. MODEL Z=1/(1 TO 8)NOISE. 01. 5

WORKSPACE COMPRESS. 04. 6

ESTIM MODMA1. METHOD EXACT. HOLD RESIDUALS(RES1). 00. 7

WORKSPACE COMPRESS. 01. 8

C -----

C EXAMEN DE LOS RESIDUALES

C 01. 9

ACF RES1. MAXLAG 20. 01. 10

WORKSPACE COMPRESS. 01. 11

C -----

C DETECCION DE OBSERVACIONES ATIPICAS DE LA SERIE

C CONTAMINADA

C 01. 12

OUTLIER MODMA1. STOP CRITICAL(3.5). TYPES AO IO LS.

WORKSPACE COMPRESS.

C ESTIMACION DE LOS EFECTOS DE LAS INTERVENCIONES

C

TSMODEL MODMA2. MODEL Z= (NWA0)INT125AO(BINARY)+ @
 (NWLS)INT50LS(BINARY)+ @
 (NWIO)/(1-P1*B-P2*B**2-P3*B**3-P4*B**4- @
 P5*B**5-P6*B**6-P7*B**7-P8*B**8) @
 INT90IO(BINARY)+ @
 1/(1-P1*B-P2*B**2-P3*B**3-P4*B**4- @
 P5*B**5-P6*B**6-P7*B**7-P8*B**8)NOISE.

ESTIM MODMA2. METHOD EXACT. HOLD RESIDUALS(RES2) FITTED(PRED).
 WORKSPACE COMPRESS.

C -----

C EXAMEN DE LOS RESIDUALES

C

ACF RES2. MAXLAG 20.
 WORKSPACE COMPRESS.

C -----

C FILTRACION DE LA SERIE ORIGINAL POR LAS

C OBSERVACIONES ATIPICAS DETECTADAS

C

TSMODEL MODMA3. MODEL Z= (NWA0)INT125AO(BINARY)+ @
 (NWLS)INT50LS(BINARY)+ @
 (NWIO)/(1-P1*B-P2*B**2-P3*B**3-P4*B**4- @
 P5*B**5-P6*B**6-P7*B**7-P8*B**8) @
 INT90IO(BINARY)+NOISE.

WORKSPACE COMPRESS.

FILTER OLD Z. NEW ZF. MODEL MODMA3.

WORKSPACE COMPRESS.

```
C      IDENTIFICACION DEL MODELO ORIGINAL
C
IDEN ZF. MAXLAG 20.
WORKSPACE COMPRESS.
C      ESTIMACION DEL MODELO FINAL
F=-1
TSMODEL MODMA4. MODEL Z= (NWA0)INT125AO(BINARY)+ @
                        (NWLS)INT50LS(BINARY)+ @
                        (F)(NWIO)(F+THETA1*B)INT90IO(BINARY)+ @
                        (1-THETA1*B)NOISE. FIXED F.
WORKSPACE COMPRESS.
C -----
C      EXAMEN DE LOS RESIDUALES
C
IDEN RES. MAXLAG 36.
WORKSPACE COMPRESS.
RETURN
```

Referencias

- Box, G.P.E. and Jenkins, G.M. -1970-. *Time Series Analysis: Forecasting and Control*. San Francisco: Hoden-Day. -Edición revisada en 1976-.
- Chang, I., Tiao G.C. and Chen, C. -1988-. *Estimation of Time Series Parameters in the Presence of Outliers*. *Technometrics* 30:193-204.
- Chen, C, and Liu, L-M -1990-. *Joint Estimation of Model Parameters and Outlier Effect in Time Series*. Working Papers Series, Scientific Computing Associates, P.O. Box625, DeKalb, Illinois 60115.

