Psicothema, 1999. Vol. 11, nº 2, pp. 409-419
ISSN 0214 - 9915 CODEN PSOTEG

SOFTWARE, INSTRUMENTACIÓN Y METODOLOGÍA

COINTEGRACIÓN EN SERIES TEMPORALES MULTIVARIADAS

Jesús Rosel, Pilar Jara y Juan Carlos Oliver
Universidad Jaume I

El presente trabajo tiene como objetivo establecer una ecuación de regresión con tres variables temporales no estacionarias. Se pretende comprobar cuál es el efecto del tamaño de una población y el índice de contaminación atmosférica sobre el número de ingresos en urgencias de enfermedades de pulmón. Se ponen a prueba varios sistemas de regresión: regresión con variables sin transformar, con variables diferenciadas, y mediante función de transferencia de Box-Jenkins, presentando todas ellas problemas de ajuste a los datos. Se comprueba cómo las tres variables están *cointegradas* y cómo existe una ecuación de regresión con un *mecanismo de corrección de error* que ajusta correctamente las variables. Además, se exponen las ventajas de la cointegración y del mecanismo de corrección de error sobre otros sistemas de regresión.

Cointegration in multivariate time series'. The aim of this paper is to establish a forecast equation with three non-stationary time series variables. An attempt will be made to show the effect of the size of a community and the level of atmospheric pollution on the number of admissions with chest illnesses to the emergency department of a hospital. Various regression systems are put to the test: regression with variables without transformation, with differentiated variables and through a Box-Jenkins transfer function, all of which show problem in fitting the data. It is found that the three variables are cointegrated and that a regression equation with an error correction mechanism which correctly adjusts the variables actually exists. The advantages of the cointegration and the error correction mechanism over other regression systems are shown.

En el pronóstico del comportamiento longitudinal mediante modelos de series temporales, se han utilizado varios procedimientos, entre los cuales destacan: *a)* los modelos *ARIMA* de Box y Jenkins (Box & Jenkins, 1970), *b)* los sistemas de análisis mediante '*vector autorregresivo* (*VAR*)' (Hannan, 1970), y *c)* los procedimientos de comprobación de hipótesis mediante *modelado dinámico* (*MD*) (Sargan, 1964; Hendry, 1979; Gilbert, 1986).

Tanto el *ARIMA* como el *VAR* son sistemas de ajuste de datos de tipo *ateórico*. De

Correspondencia: Jesús Rosel
Departamento de Psicología, E., E., S. y Metodología
Universitat Jaume I
12080 Castellón (Spain)
E-mail: rosel@psi.uji.es

cualquier forma, el modelo *ARIMA* sigue siendo muy válido en modelos univariados y en el estudio del impacto de un tratamiento (para ver el tipo de efecto temporal producido en la variable). No obstante, desde una perspectiva teórica es más válido (sobre todo para modelos multivariados) el sistema de *MD*. Los procedimientos desarrollados mediante *MD* tienen la ventaja de que han de ser estimados mediante una *hipótesis teórica*, para después comprobar si esa hipótesis inicial es correcta. Los partidarios del *MD* han desarrollado una metodología cuyo fundamento está basado en el de la regresión clásica (univariada o multivariada, con una sola variable dependiente o con varias, mediante sistemas de ecuaciones simultáneas con variables retardadas o sin ellas, con lo cual estos modelos se relacionan con los de ecuaciones estructurales).

El problema nuclear del investigador que trabaja con datos temporales es el de establecer el 'proceso generador de datos' que ha dado lugar a la serie temporal real objeto de estudio (y cuya observación responde a una determinada 'realización'), pero ese proceso generador de datos ha de estar provocado por un mecanismo que debe responder a supuestos substantivos sobre cómo se han producido los datos. Es decir, una buena teoría ha de explicar los datos (comprobada mediante el correspondiente modelo estadístico), y no puede dejarse a un modelo estadístico (por sí solo) que substituya a la teoría.

Mediante el enfoque de *MD*, la teoría y el ajuste estadístico se han de complementar, en lugar de plantearse cada uno de ellos aisladamente. Uno de los principales méritos de los teóricos del *MD* es el de haber establecido mecanismos de pronóstico de las series a largo plazo (además de a corto plazo), pero sobre todo, han desarrollado un conjunto de procedimientos (tests estadísticos) para comprobar que los supuestos hipotéticos del modelo planteado se cumplen correctamente.

El objetivo general del presente trabajo es presentar algunos de los avances de la modelización estadística mediante modelado dinámico, y principalmente la prueba de *estacionalidad* o de *raíces unitarias* de Dickey-Fuller (1979), la de *cointegración* de Johansen (1988) y el '*mecanismo de corrección de error*, o: *MCE'* (Phillips, 1957; Sargan, 1964). Para llevar a cabo el objetivo inicial, se intentarán exponer las diferentes técnicas estadísticas de la manera más asequible, orientando al lector sobre aspectos más técnicos en la bibliografía que se cita.

Se parte del supuesto de que en una gran ciudad en vías de desarrollo (en la que hay una gran inmigración y en la que se instalan industrias contaminantes), el número de personas que son atendidas de enfermedades de pulmón en los servicios de urgencias de las unidades hospitalarias de esa ciudad durante un mes cualquiera, es función del número de habitantes de esa ciudad, y del índice de contaminación de la atmósfera (medido en la cantidad de dióxido de azufre, expresado en $\mu g/m^3$ N día, es decir, en microgramos por metro cúbico bajo condiciones normales). Se tienen las siguientes variables temporales (tomándose los datos de cada variable mes a mes): el número de personas atendidas en urgencias de pulmón en los hospitales de la ciudad (UP_t), el número total de personas que habitan en esa ciudad (H_t), y el índice medio de contaminación atmosférica por dióxido de azufre en un observatorio del centro de la ciudad (DA_t).

Para desarrollar el objetivo inicial nos valdremos de datos coherentes con el problema, simulados por los autores; la variable H_t, se ha generado mediante el proceso: $\nabla H_t = C_1 + 0.3\nabla H_{t-1} + a_{1,t}$; mientras la variable DA_t sigue un proceso: $DA_t = C_2 + DA_{t-1} + 0.75 + a_{2,t}$, en las anteriores ecuaciones, $a_{1,t}$ y $a_{2,t}$ son procesos de ruido blanco, y C_1, C_2 constantes. UP_t se genera a partir de H_t y de DA_t del siguiente modo: $UP_t = -9.000 + 0.0025 \cdot H_t + 2.5 \cdot DA_t + a_t$. En la Tabla 1 se exponen los datos de las variables.

Tabla 1
Datos de las variables *t* (mes de medición), UP_t, H_t y DA_t

t	UP_t	H_t	DA_t	*t*	UP_t	H_t	DA_t	*t*	UP_t	H_t	DA_t
1	5440	6069238	10,3	*37*	5625	6074467	57,8	*73*	5721	6104831	71,1
2	5451	6070294	11,5	*38*	5612	6074254	56,3	*74*	5727	6104031	65,4
3	5465	6071867	23,2	*39*	5600	6076828	62,4	*75*	5722	6104146	68,8
4	5492	6073466	21,4	*40*	5599	6076332	59,6	*76*	5711	6102971	75,2
5	5516	6075107	29,7	*41*	5599	6074427	61,0	*77*	5708	6105053	79,3
6	5544	6079032	32,3	*42*	5598	6075663	60,4	*78*	5714	6107390	79,3
7	5569	6078767	32,0	*43*	5603	6076582	56,6	*79*	5725	6103204	88,2
8	5579	6080038	26,2	*44*	5603	6077538	54,8	*80*	5734	6099075	87,1
9	5574	6079109	30,0	*45*	5603	6076834	57,3	*81*	5735	6097475	83,4
10	5558	6080175	31,2	*46*	5601	6078760	59,0	*82*	5726	6095175	83,9
11	5545	6078366	30,0	*47*	5605	6082220	61,2	*83*	5715	6093690	87,3
12	5535	6079865	26,7	*48*	5622	6087603	63,5	*84*	5708	6092812	84,0
13	5527	6078160	35,1	*49*	5648	6089118	67,1	*85*	5704	6091663	82,2
14	5526	6079021	39,5	*50*	5673	6090652	66,8	*86*	5697	6093386	81,8
15	5539	6078622	40,4	*51*	5692	6087949	66,3	*87*	5697	6094580	85,1
16	5555	6080562	37,9	*52*	5691	6090134	63,5	*88*	5704	6095649	86,1
17	5569	6082693	39,7	*53*	5684	6089912	73,9	*89*	5717	6098315	90,1
18	5584	6082731	43,1	*54*	5677	6092625	67,9	*90*	5730	6100226	86,0
19	5594	6084483	49,9	*55*	5673	6094082	72,7	*91*	5743	6100954	93,1
20	5611	6083775	52,3	*56*	5670	6093948	71,3	*92*	5755	6100354	91,0
21	5625	6084057	56,0	*57*	5674	6096867	76,8	*93*	5763	6099314	93,4
22	5638	6084833	57,3	*58*	5686	6099677	83,8	*94*	5766	6098568	94,0
23	5644	6083800	53,2	*59*	5712	6100802	80,9	*95*	5765	6099272	94,8
24	5638	6083037	51,8	*60*	5735	6103095	78,8	*96*	5764	6098905	90,4
25	5624	6081245	56,6	*61*	5747	6103855	75,9	*97*	5753	6096251	90,9
26	5610	6080246	52,5	*62*	5747	6107862	71,1	*98*	5735	6093933	101,3
27	5595	6079430	55,4	*63*	5738	6107930	66,7	*99*	5731	6094252	101,2
28	5588	6080837	51,4	*64*	5723	6107824	66,8	*100*	5739	6093879	97,8
29	5588	6079292	44,1	*65*	5705	6108719	62,5	*101*	5741	6091252	93,7
30	5576	6078442	49,9	*66*	5688	6106773	57,6	*102*	5737	6093412	91,7
31	5575	6077077	57,4	*67*	5667	6105853	58,0	*103*	5733	6093884	98,3
32	5586	6075278	61,0	*68*	5652	6103585	58,4	*104*	5738	6095933	87,7
33	5599	6076043	63,1	*69*	5641	6104316	67,2	*105*	5741	6095892	94,3
34	5618	6075873	61,7	*70*	5649	6106044	66,5	*106*	5739	6096420	87,6
35	5629	6074754	64,4	*71*	5668	6107558	70,7				
36	5632	6075210	57,8	*72*	5694	6108265	74,8				

Es decir, la hipótesis substantiva sería: $UP_t = f(H_t, DA_t)$, que daría lugar a la hipótesis estadística:

$$UP_t = b_0 + b_1 H_t + b_2 DA_t + a_t \qquad (1)$$

En la ecuación *(1)*, b_0, b_1 y b_2 son los coeficientes (que han de ser estimados), y a_t es un proceso de *ruido blanco*.

Análisis preliminares y regresión

Se ha llevado a cabo un análisis del número de raíces unitarias de cada serie mediante el test de *Dickey-Fuller aumentado* (*DFA*), comprobándose que para la serie UP_t, el *DFA(1)* da un valor de $t=-1.691$ (*n.s.*) mientras el *DFA(2)* da un valor de $t=-3.4542$ ($p<0.05$); para la serie H_t, el *DFA(1)*, da una $t=-0.0233$ (*n.s.*), el *DFA(2)*$=-3.8776$ ($p<0.05$); y para la variable DA_t, el DFA(1)

da *t=-1.652* (*n.s.*) siendo el *t=-5.3537* (*p<0.05*) del *DFA(2)*. Las tres series son integradas de primer orden (UP_t *~I(1),* H_t *~I(1),* DA_t *~I (1)*), lo cual implica que cada serie ha de ser autodiferenciada una vez para que sea estacionaria (Dickey and Fuller, 1979, 1981; Phillips and Perron, 1988; MacKinnon, 1991).

Debido a la hipótesis establecida en la relación de la ecuación *(1)*, consiguiéndose los estimadores de los coeficientes mediante procedimiento de mínimos cuadrados ordinarios (MCO), puesto que (como el teorema de Gauss-Markov demuestra) el sistema de MCO es el *mejor estimador lineal insesgado* bajo condiciones de normalidad y no autocorrelación de los residuales (Dunteman, 1984). En este trabajo, los estimadores se han calculado mediante MCO (salvo que se indique expresamente). Se obtienen los siguientes resultados:

$$UP_t = \underset{(-6.61)}{-9887.1} + \underset{(10.25)}{0.0025 \cdot H_t} + \underset{(18.81)}{2.493 \cdot DA_t} + a_t \quad (2)$$

En la ecuación *(2)*, y en sucesivas ecuaciones, se representa inmediatamente debajo de los coeficientes los respectivos valores del estadístico *t* de *Student* correspondientes a cada uno de ellos. Se comprueba que los valores de cada *t* son muy significativos. Pero si atendemos al valor de otros estadísticos, se obtienen los siguientes resultados: $R^2 = 0.927$ (*p<0.000*), *estadístico de Durbin-Watson* de los residuales (a partir de ahora: *DW*)*=0.579* (*p<0.000*). En la Figura 1 se representan los valores de los residuales de pronóstico de la ecuación *(2)*.

A la vista de los resultados de la ecuación *(2)*, se concluye que los residuales están autocorrelacionados, y por lo tanto, dicha ecuación no es correcta, encontrándonos ante un caso claro de *correlación espuria*, fundamentalmente porque los estimadores de los errores estándar son inconsistentes (Yule, 1926; Granger & Newbold, 1974; Banerjee *et alt*., 1993).

Teniendo en cuenta lo anterior, se podría intentar establecer un sistema de modelización de los residuales, completando la anterior ecuación *(2)*, haciendo que cada error sea ahora la variable dependiente hasta dejar los residuales sin autocorrelación, utilizando, p.ej., cualquiera de los siguientes procedimientos: el de Cochrane-Orcutt, el de Hildreth-Lu o el de las primeras diferencias (Neter, Wasserman & Kutner, 1990). No obstante, los anteriores procedimientos son formalmente correctos pero substantivamente improcedentes por dos motivos: *a)* no tiene un significado real (una referencia con la realidad) el ajuste de los residuales, y *b)* no responden a los parámetros de la hipótesis inicial (recuérdese que la hipótesis y el sistema generador de la función que relaciona las tres variables es: $UP_t = f(H_t, DA_t)$).

Una alternativa (aunque incorrecta, puesto que tan sólo establecería la dinámica a corto plazo), al ser las tres series integradas de primer orden (Y_t*~I(1)*), sería realizar una regresión de las puntuaciones autodiferenciadas de primer orden (la autodiferenciación de una serie es una nueva serie temporal en la que a cada valor se le resta el inmediatamente anterior), con lo cual, una reparametrización de la misma hipótesis *(1)* sería (Fuller, 1976):

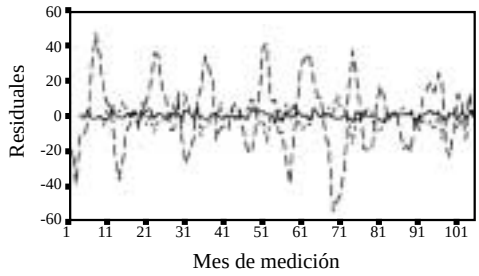

Figura 1. Gráfico de los residuales de UP_t *para cada mes de medición:*
a) de la ecuación de regresión ordinaria (2), identificados como a_t, _ _ _ _ _ _ _ _ _
b) de los residuales de la función de transferencia Box-Jenkins (5), identificados por e_t, _ _ _ _ _ _ _ _ _ _
c) de la ecuación de cointegración (13), identificados como u_t. _____

$$\nabla UP_t = f(\nabla H_t, \nabla DA_t) \qquad (3)$$

porque cada serie diferenciada es la misma serie a la que se le resta el valor anterior de la misma serie retardada. El símbolo ∇ representa el operador de retardos; así, dada una serie temporal, Y_t, si se le aplica a cada valor el operador de diferenciación: $\nabla Y_t = (1\text{-}B) = Y_t - Y_{t-1}$; en general: $B^k Y_t = Y_{t-k}$. La estimación de los coeficientes correspondientes a la ecuación (3) es:

$$\nabla UP_t = \underset{(1.92)}{2.3055} - \underset{(2.84)}{0.002}\cdot\nabla H_t + \underset{(0.22)}{0.061}\cdot\nabla DA_t + u_t, \qquad (4)$$

los estadísticos de conjunto de esta ecuación son: $R^2 = 0.073$ (p=0.020), y el estadístico de $DW=0.583$ (p=0.000). Por lo que esta ecuación tampoco es aceptable porque si bien R^2 es significativo, los residuales están autocorrelacionados; nótese cómo el coeficiente de ∇DA_t en (4) es no significativo, pero el proceso generador de ∇UP_t incluye la variable ∇DA_t. Llegados a este punto, una solución podría ser la de llevar a cabo ecuaciones autorregresivas de la fórmula *(4)*, así, plantear: $\nabla UP_t = f(\nabla H_t, \nabla H_{t-1}, \nabla DA_t, \nabla DA_{t-1})$, y tal vez se consiguiese ajustar la serie.

Otra solución podría ser la de establecer la función de transferencia de (UP_t), en función de H_t y de DA_t, mediante la búsqueda del modelo de función de transferencia de Box-Jenkins (1970) que ajuste los datos; después de haber realizado las correspondientes comprobaciones: diagnóstico, identificación, estimación y comprobación; mediante procedimiento de *preblanqueo* y por estimación mediante el método *backcasting*, el modelo propuesto es:

$$UP_t = 5652.62 - (5.65.10^{-6})H_{t-1} + (0.56 + 0.38B^1 - 0.34B^6)DA_{t-1} + \frac{(1-045.B)}{(1-B)(1-0.87B)}a_t \qquad (5)$$

Comprobándose que $R^2 = 0.0998$ y que el estadístico $DW=2.012$; es decir, los residuales son 'ruido blanco'. Se podría haber especificado cómo se ha llegado a la formulación *(5)*, pero no es objeto del presente trabajo, dejándose al lector que haga la correspondiente estimación y la interpretación de esta ecuación.

En cualquiera de los anteriores casos *(3)*, *(4)* y *(5)* se está forzando el hallazgo de una solución que no se corresponde con la hipótesis de partida. La solución más correcta consiste en comprobar si las series originales (UP_t, H_t y DA_t) están *cointegradas*, con el fin de aplicar una ecuación con *mecanismo de corrección de error*, que siendo equivalente a la ecuación *(1)* (es decir, una *reparametrización* de esa misma ecuación) cumpla con los requisitos estadísticos adecuados.

Cointegración y mecanismo de corrección de error

Supóngase dos variables temporales integradas de orden d (sean $X_t \sim I(d)$, e $Y_t \sim I(d)$), si se realiza una combinación lineal entre ambas, lo más probable es que dicha combinación sea también $I(d)$; es decir, si se tiene en cuenta que la regresión entre ambas ($Y_t = b_0 + b_1 X_t + e_t$) es un caso especial de combinación lineal entre X_t e Y_t, los residuales, e_t (siendo: $e_t = Y_t - b_0 - b_1 X_t$), también serán $I(d)$. Ahora bien, si existe un coeficiente en la anterior ecuación que cumpla el requisito: $e_t \sim I(0)$, se dice que ambas series son *cointegradas completas* de orden d (en forma compacta, se expresa: $X_t, Y_t \sim CI(d,d)$) (Sargan, 1964; Davidson *et alt*., 1978; Granger, 1981; Engle & Granger, 1987).

Es muy importante detectar que varias series están cointegradas, porque indica que dichas series pueden admitir una formulación en la que sus *residuales dejan un ruido blanco* (el analista de datos ha de saber conseguir la formulación adecuada, que en ocasiones puede ser simple, aunque a veces se

ha de hacer una reparametrización de la misma). El concepto de cointegración remite al de valor esperado de las series a *largo plazo*, pues si bien cada una de ellas por separado muestra tendencia, las diferencias entre ellas (en función de sus respectivos coeficientes) tiende a ser constante (Pesaran, 1987).

En el caso que nos ocupa, al ser UP_t, H_t y $DA_t \sim I(1)$ ya se ha visto en la ecuación *(2)* que los residuales son también $e_t \sim I(1)$. La cuestión clave es comprobar si existe un vector de coeficientes que afecte a UP_t, H_t y DA_t y que deje unos residuales de orden *cero* ($e_t \sim I(0)$), en el caso de que esto ocurriese, la ecuación establecida conforme a ese vector de coeficientes sería correcta, y UP_t, H_t y DA_t estarían cointegradas (UP_t, H_t y $DA_t \sim CI(1,1)$).

Los sistemas más utilizados para llevar a cabo un procedimiento de regresión de series temporales múltiples mediante cointegración son: a) el de Johansen (1988, 1992), quien ha desarrollado un método que permite comprobar si existe el posible vector (o posibles vectores) de cointegración, y cuál (o cuáles) de esos vectores es significativo, y b) el de Engle y Granger (1987), que requiere un álgebra más sencilla, y expondremos a continuación.

El procedimiento de Engle y Granger supone la comprobación de varios pasos, en el caso de dos variables (X_t e Y_t), se procedería:

a) Hallar el orden de integración de las variables del sistema, si las dos son de distinto orden, no puede haber relación lineal entre ambas (no cointegran). Si las dos son del mismo orden, tal vez cointegren; pues que sean integradas del mismo orden es una condición necesaria, pero no suficiente, para la cointegración.

b) Estimar la ecuación de regresión que responda a la hipótesis (normalmente: $Y_t = ß_0 + ß_1 X_t + e_t$), guardando los residuales de esta ecuación, que equivalen a: $e_t = Y_t - ß_0 - ß_1 X_t$ (6)

c) Comprobar que los residuales e_t calculados en (6) son estacionarios,mediante, p.ej., el test de Dickey y Fuller. Si los residuales son estacionarios, es señal de que $e_t \sim I(0)$, y que las variables X_t, $Y_t \sim CI(d,d)$, cointegran totalmente.

d) Si se cumple la anterior condición *c)*, se puede establecer la ecuación que contiene el mecanismo de corrección de error:

$$\phi(B)\cdot \nabla Y_t = \varphi(B)\cdot \nabla X_t + \eta\,(Y_{t-p} - ß_0 - ß_1 X_{t-p}) + u_t \qquad (7)$$

en esta ecuación, $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_a B^a$, es decir, se trata de un polinomio autorregresivo de orden *a*; $\varphi(B) = 1 + \varphi_1 B + \varphi_2 B^2 + \ldots + \varphi_b B^b$, siendo un polinomio de operador de retardos de orden *b*. Obsérvese cómo en la ecuación *(7)* el término que aparece entre paréntesis es el residual (o el *error*) de la ecuación *(6)* retardado *p* unidades de tiempo.

De acuerdo con el modelo general de *MCE*, y aplicándolo de manera simplificada al problema que estamos tratando, una reparametrización de la ecuación *(1)* conforme a la ecuación (7) de MCE, sería:

$$\phi(B)\cdot \nabla UP_t = \varphi(B)\cdot \nabla H_t + \gamma(B)\cdot \nabla DA_t + \eta\,(UP_{t-1} - b_0 - b_1 H_{t-1} - b_2 DA_{t-1}) + u_t \qquad (8)$$

Obsérvese cómo en la ecuación *(8)* el término que aparece entre paréntesis es el residual (o el *error*) de la ecuación *(2)* retardado una unidad de tiempo.

Se ha llevado a cabo la estimación de los coeficientes de la ecuación *(8)* mediante el procedimiento de estimación no-lineal con el algoritmo de estimación de Levenberg-Marquardt (Norusis, 1993); se utiliza la estimación no-lineal puesto que hay coeficientes introducidos en un paréntesis que multiplican a su vez a η; y una vez eliminados los coeficientes de las variables diferenciadas no significativos queda la ecuación:

$$\nabla UP_t = \underset{(9.88)}{0.493}\cdot \nabla UP_{t-1} + \underset{(3.85)}{0.207}\cdot \nabla UP_{t-2} + \underset{(9.55)}{0.001}\cdot \nabla H_{t-1} + \underset{(2.92)}{0.212}\cdot \nabla DA_{t-1} \qquad (9)$$

$$+ 0.386 \cdot \nabla DA_{t-2} - 0.305\ (UP_{t-1} + 8009.1 - 0.002 \cdot H_{t-1} - 2.590 \cdot DA_{t-1}) + u_t \quad (9)$$
$$(6.74) \qquad (13.26) \qquad (17.27) \quad (28.98) \quad (60.00)$$

donde el término entre paréntesis es el *MCE*. En esta ecuación: $R^2 = 0.974$ ($p < 0.000$), *estadístico* $DW = 2.087$ ($p = 0.451$). Es decir, la ecuación es estadísticamente correcta, porque tiene todos los coeficientes significativos y los residuales son ruido blanco (más adelante se explica por qué el coeficiente del *MCE*, η, ha de tener signo negativo). Se ha llevado a cabo el test estadístico *t* para comparar las varianzas de residuales relacionados (Amón, 1994) de la ecuación de cointegración (*9*) con los de la función de transferencia (*5*), dando un valor de $t = 2.71$ ($p < 0.01$), lo cual indica que los residuales de (*9*) son significativamente menores que los de (*5*), al ser su varianza menor, con lo cual puede inferirse que ajusta significativamente mejor el modelo de MCE que el modelo de Box-Jenkins.

En la Figura 1 se representan los valores de los residuales de pronóstico de la ecuación *(2)*, los de la *(5)* y los de la *(9)*, con el fin de comparar los respectivos valores. Se aprecia cómo los valores de los residuales de la ecuación *(9)* son significativamente menores, además de no estar autocorrelacionados.

Ante estos resultados surge la cuestión (Pesaran & Pesaran, 1992) de qué relación tienen la ecuación *(2)* y la *(9);* la respuesta a esta cuestión se puede resolver pensando en la *solución a largo plazo*, en la que hubiese un *estado de equilibrio* (se entiende por *estado de equilibrio* la situación en la cual no hay tendencia al cambio), se daría cuando:

$$\ldots = UP_{t-1} = UP_t = UP_{t+1} = \ldots = UP^*, \text{ por tanto: } \nabla UP_t = 0 \quad (10)$$
$$\ldots = H_{t-1} = H_t = H_{t+1} = \ldots = H^*, \text{ por tanto: } \nabla H_t = 0,$$
$$\ldots = DA_{t-1} = DA_t = DA_{t+1} = \ldots = DA^*, \text{ por tanto: } \nabla DA_t = 0,$$

Así, la relación de los valores de *(8)* a largo plazo sería:

$$0 = -0.305\ (UP^* + 8009.1 - 0.002 \cdot H^* - 2.590 \cdot DA^*), \quad (11)$$
$$0 = UP^* + 8009.1 - 0.002 \cdot H^* - 2.590 \cdot DA^*,$$
$$UP^* = -8009.1 + 0.002 \cdot H^* + 2.590 \cdot DA^*$$

Estos valores coinciden sensiblemente con los de la regresión estándar (ver ecuación *(2)*), la única diferencia radica en el nivel de la ecuación (la constante), pero ha de considerarse que la constante en un proceso a largo plazo no es fidedigna, puesto que las series no son estacionarias (por tanto su nivel varía con el tiempo, y también variaría la constante en distintos intervalos de tiempo). La gran ventaja de la ecuación *(9)* sobre las otras (aparte de su adecuación y de su significación estadística) es que el *MCE* ajusta los valores de la tendencia de la serie a largo plazo, mientras los términos con los valores diferenciados de las variables ajustan la serie a corto plazo.

Se puede llegar a la ecuación *(9)* mediante el sistema estándar de regresión si no se dispone de un programa con un sistema de estimación no-lineal, para ello se ha de hacer la estimación de Engle y Granger; así, en el paso *b)* del sistema de estimación de Engle y Granger se ha de efectuar el cálculo de los coeficientes de la ecuación de regresión de (*6*) mediante MCO, guardándose los errores de esta ecuación retardada: $e_t = Y_t - ß_0 - ß_1 X_t$; a continuación, en el paso *d)* de Engle y Granger se introducen los errores de pronóstico (retardados una unidad temporal) en el MCE de la fórmula *(8):*

$$\phi(B) \cdot \nabla UP_t = \varphi(B) \cdot \nabla H_t + \varphi(B) \cdot \nabla DA_t + \eta\,(e_{t-1}) + u_t \quad (12)$$

se observa que en las ecuaciones *(8)* y *(12)* los términos entre paréntesis (el *MCE*) son equivalentes. En nuestros datos, a partir de la ecuación *(2)*:

$$UP_{t-1} + 9887.1 - 0.0025.H_{t-1} - 2.493.\ DA_{t-1} = a_{t-1} \quad (13)$$

Si en *(12)* se inserta el residual a_{t-1}de (*13*) en el paréntesis del *MCE*:

$$\phi(B)\cdot \nabla UP_t = \varphi(B)\cdot \nabla H_t + \gamma(B)\cdot \nabla DA_t + \eta(a_{t-1}) + u_t \quad (14)$$

Se ha efectuado una regresión por MCO según el modelo de la ecuación *(14)*, dando los siguientes coeficientes:

$$\nabla UP_t = \underset{(10.10)}{0.527}\cdot \nabla UP_{t-1} + \underset{(2.97)}{0.167}\cdot \nabla UP_{t-2} + \underset{(8.99)}{0.001}\cdot \nabla H_{t-1} + \underset{(3.84)}{0.277}\cdot \nabla DA_{t-1}$$

$$+\underset{(6.73)}{0.400}\cdot \nabla DA_{t-2} - \underset{(12.18)}{0.286}(a_{t-1}) + u_t \quad (15)$$

Siendo el valor $R^2 = 0.973$ ($p<0.000$), y el *estadístico DW* de los residuales $=1.887$ ($p=0.562$). Los valores de los coeficientes de esta ecuación son sensiblemente iguales a los de la *(9)*, pero los residuales de la ecuación *(15)* son más correctos que las de la *(9)*, porque la *(15)* tiene más grados de libertad al tener que estimar un menor número de coeficientes.

La ecuación (8) también se podía haber parametrizado del siguiente modo: $\eta b_0 = \eta_0$, $\eta b_1 = \eta_1$, $\eta b_2 = \eta_2$, con lo cual quedaría:

$$\phi(B)\cdot \nabla UP_t = \varphi(B)\cdot \nabla H_t + \gamma(B)\cdot \nabla DA_t + \eta UP_{t-1} - \eta_0 - \eta_1 H_{t-1} - \eta_2 DA_{t-1} + u_t \quad (16)$$

de esta forma, la estimación de los coeficientes puede hacerse linealmente, pero es preferible el procedimiento de estimación no lineal (*8*) o el de la substitución del residual (*12*) porque así se tiene el sentido del MCE como residual de momento anterior que entra en la regresión, además de que en muestras pequeñas pueden variar los valores de estimación de los coeficientes.

Discusión

A lo largo de estas páginas hemos intentado realizar una introducción a las propiedades de la cointegración, vinculando el modelo de cointegración con el de *MCE*. Se han generado unos datos temporales mediante simulación, y se ha comprobado su ajuste mediante distintos procedimientos. Se ha seguido el sistema de Engle y Granger (1987) para estimar la ecuación de ajuste de los datos, se comprueba: *a)* que las distintas series son diferenciables de orden uno (*I(1)*), *b)* son modelizables mediante mecanismo de corrección de error.

Por medio del *MCE* se ha comprobado: *a)* que cumple con la hipótesis inicial y con el mecanismo generador de los datos, *b)* que cumple con las condiciones de equilibrio de la serie a largo plazo, *c)* el hecho de incluir las variables diferenciadas en la ecuación de regresión de *MCE* contribuye al ajuste del pronóstico a corto plazo de la serie que es variable dependiente, y *d)* el *MCE* constituye el sistema de ajuste de la serie a largo plazo. La equivalencia entre las ecuaciones (*2*) y (*9*) viene dada por el hecho de que la ecuación (*9*) es una reparametrización de la (*2*) en el componente a largo plazo.

El *MCE* fue propuesto y utilizado originariamente por Philips (1957) y Sargan (1964), si bien fue demostrada su generalización por Engle y Granger (1987). Teóricamente, existe una redundancia perfecta entre la significación del coeficiente del MCE, la significación del test de cointegración y la regresión ordinaria cuando la muestra de datos es infinita, pero ha de comprobarse en cada respectiva muestra de datos la significación de cada uno de ellos porque a veces no presentan redundancia perfecta.

En cualquier caso, se ha de seguir una serie de pasos para la modelización de un proceso multivariado de series temporales mediante un sistema de regresión: *a)* establecimiento de una hipótesis substantiva, *b)* cálculo de los correspondientes tests de raíces unitarias para cada variable por separado (con el fin de comprobar que tienen el mismo orden de integración), y *c)* ajuste mediante sistema de regresión no-lineal del *MCE* coherente con la hipótesis (el que las series tengan el mismo orden de integración no garantiza el que estén cointegradas, pues

que las series sean del mismo orden es una condición necesaria, pero no suficiente, para la modelización estadística).

Entre las implicaciones del uso de *MCE* destacan que en una ecuación de regresión no puede incluirse una variable dependiente que sea de mayor orden de integración que cada una de las variables independientes, puesto que siempre el residual sería de un orden igual o superior a la unidad (cabría establecer otras estrategias: p.ej., diferenciar las series hasta que fuesen de orden cero, pero se perdería la relación dinámica de las series a largo plazo; una alternativa más correcta para captar la dinámica a largo plazo sería la de integrar las series de menor orden hasta alcanzar el orden de la serie de mayor orden de diferenciación).

Una ecuación de regresión con MCE consta de dos partes: el término de *MCE* retardado (con lo cual se respeta el sentido de la hipótesis original, a la vez que se refleja en él la *dinámica a largo plazo*), y los términos diferenciados, que reflejan la *dinámica transitoria* del sistema, puesto que tan sólo proporcionan el ajuste temporal a corto plazo del modelo. Así, p.ej.: en la ecuación (*9*), el MCE estaría incluido en la expresión: $(UP_{t-1} + 8009.1 - 0.002 \cdot H_{t-1} - 2.590 \cdot DA_{t-1})$, mientras los términos diferenciados: $0.493 \cdot \nabla UP_{t-1} + 0.207 \cdot \nabla UP_{t-2} + 0.001 \cdot \nabla H_{t-1} + 0.212 \cdot \nabla DA_{t-1} + 0.386 \cdot \nabla DA_{t-2}$ ajustan la serie de la variable dependiente a corto plazo.

El significado intuitivo del término de *MCE* es que, por un lado, el error de la ecuación *(2)* se convierte en variable de pronóstico de la ecuación *(9)*, con lo que la varianza de los errores de pronóstico de esta ecuación ha de ser menor que la de la *(3);* pero por otro lado, el término de *MCE* se convierte en *atractor* de equilibrio de la ecuación, pues al ser el signo del coeficiente negativo, cuando el error en el momento *t-1* es positivo influye negativamente en el momento *t* y viceversa, haciendo que los errores de la nueva ecuación se distribuyan alrededor del valor *cero*. En la ecuación *(9)* se observa que los signos del *MCE* son consistentes con la hipótesis (el número de habitantes y la contaminación influyen positivamente sobre el número de urgencias de pulmón) y que el signo del coeficiente de todo el *MCE* es igualmente consistente (es negativo).

Se ha llevado a cabo una regresión de la variable dependiente mediante una ecuación de transferencia por el procedimiento de Box-Jenkins, se ha comprobado que los residuales del MCE tienen menor varianza que los de Box-Jenkins. La ventaja, para el analista, del procedimiento de *MCE* sobre el modelado *ARIMA* es que resulta más cómodo (además de substantivamente más correcto) calcular una sola ecuación de regresión estándar (como se hace en el *MCE*). Los modelos *ARIMA*, utilizados ateóricamente, presentan el inconveniente de que pueden presentar coeficientes significativos sin serlo en la realidad, con lo que se cometerían errores de estimación tipo I. Es fácil incurrir en un error de tipo I, sobre todo, cuando se modelizan series temporales no estacionarias en media, puesto que las series presentarán correlaciones espurias por efecto de la monotonía matemática (ascendiente o descendiente) de los datos. Este fenómeno es conocido como el 'arañado de los datos' por investigadores vinculados a la Universidad de Londres (Pagan, 1990).

En cualquier caso, no ha sido nuestra intención exponer qué procedimiento es mejor o peor para establecer una correcta ecuación de regresión con variables temporales (el modelado *ARIMA* es muy válido para modelos univariados, para comprobar el efecto de un tratamiento o para detectar la presencia de valores atípicos), sino describir cómo existe un método, el de *cointegración-mecanismo de corrección de error* vinculado directamente al de la regresión clásica, que optimiza el pronóstico (la varianza

explicada) de la variable dependiente, y que está relacionado con otros procedimientos de regresión múltiple (vector autorregresivo, regresión no-lineal, etc.), mostrándose todos ellos como variantes de un mismo modelo lineal general. La gran ventaja del *MCE* es que resulta más simple en su formulación e interpretación (aunque pueda parecer más complejo a primera vista) que otros sistemas, además de que su expresión ajusta mejor al mecanismo generador de datos y responde a una hipótesis substantiva de partida.

Agradecimientos

Los autores agradecen la colaboración del profesor Dr. Vicente Esteve en relación con la información sobre aspectos técnicos en contaminación ambiental.

Este trabajo ha sido financiado con la ayuda del Fondo de Investigaciones Sanitarias, Ministerio de Sanidad y Seguridad Social (Proyecto de Investigación 97/2121).

Referencias

Amón, J. (1994) *Estadística para psicólogos* (Vol. 2). Madrid: Pirámide.

Banerjee, A., Dolado, J.J., Galbraith, J.W. y Hendry, D.F. (1993) *Co-integration, error correction and the econometric analysis of non-stationary data*. Oxford: Oxford University Press.

Box , G. E. P. y Jenkins, P. M. (1970) *Time series analysis: forecasting and control*. San Francisco: Holden-Day.

Davidson J., Hendry, D. F., Srba, F. y Yeo S. (1978) Econometric modelling of the aggregate time series relationships between consumers expenditure and income in the United Kingdom. *Economic Journal, 88*, 661-692.

Dickey, D.A., y Fuller, W.A. (1979) Distribution on the estimators for autorregressive time series with a unit root. *Journal of the American Statistical Association, 74*, 427-431.

Dickey, D.A., y Fuller, W.A. (1981) Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica, 49*, 1.057-1.072.

Dunteman, G.H. (1984) *Introduction to linear models*. Beverly Hills, CA: Sage.

Engle, R. F. y Granger, C. W. J. (1987) Cointegration and error corrections representation, estimation and testing. *Econometrica, 55*, 251-276.

Engle, R. F. y Yoo, S. (1989) *A survey of cointegration.* San Diego: University of California.

Fuller, W. (1976) *Introduction to statistical time series*. New York: J. Wiley.

Gilbert, C.L. (1986) Professor Hendry's econometric methodology. *Oxford Bulletin of Economics and Statistics, 48*(3), 283-307.

Granger, C. W. J. (1981) Some properties of time series time series data and their use in econometric model specification. *Journal of Econometrics, 16*, 121-130.

Granger, C. W. J. (1983) *Co-integrated variables and error correcting models (Discussion paper).* San Diego: University of California.

Granger, C. W. J. y Newbold, P. (1974) Spurious regression in econometrics. *Journal of Econometrics, 2*, 111-120.

Hannan, E. J. (1970) *Multiple time series*. New York: Wiley.

Hendry, D. F. (1979) Predictive failure and econometric modelling in macroeconomics: the transactions demand for money. En: P. Ormerod (Ed.) *Modelling the economy*. London: Heinemann.

Johansen, S. (1988) Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control, 12*, 231-254.

Johansen, S. (1992) Cointegration in partial systems and the efficiency of single equations analysis. *Journal of Econometrics, 52*, 389-402.

MacKinnon, J. (1991) Critical values for cointegration tests. Pgs. 267-276 en: R.F. Engle and C.W.J. Granger (Eds.) *Long-run economic relationships*. Oxford: Oxford University Press.

Neter, J., Wasserman, W. y Kutner, M.H. (1990) *Applied linear statistical models*. Boston, MA: Irwin.

Norusis, M. J. *SPSS for Windows. Advanced statistics. Release 6.0. Chicago,* IL: SPSS Inc.

Pagan, A. R. (1990). 'Three econometric methodologies', en Granger, C. W. J. *Modelling eco-*

nomic series. Readings in econometric methodology. Oxford: Clarendon Press.

Pesaran, M. H. (1987) *The limits to rational expectations*. Oxford: Blackwell.

Pesaran, H, y Pesaran, B. (1992) *Microfit, v3.0*. Oxford: Oxford University Press.

Phillips, A.W. (1957) Stabilization policy and the time forms of lagged responses. *Economic Journal, 67*, 265-277.

Phillips, P.C.B., y Perron, P. (1988) Testing for a unit root in time series regression. *Biometrica, 75*, 335-346.

Sargan, J. D. (1964) Wages and prices in the United Kingdom: a study of econometric methodology. En: P. E. Hart y J. K. Whitaker (Eds.) *Econometric analysis for national economic planning*. London: Butterworths.

Yule, G.U., (1926) Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series. *Journal of the Royal Statistical Society, 89*, 1-64.

Aceptado el 7 de julio de 1998