Uso correcto de la correlación cruzada en Climatología: el caso de la presión atmosférica entre Taití y Darwin*

Correct use of the correlation crossed in Climatology: case of the atmospheric pressure between Taití and Darwin

José Manuel Guevara Díaz

RESUMEN

La correlación entre dos series temporales, tanto en tiempo simultáneo, como desfasadas en el tiempo, se conoce como correlación cruzada y habitualmente se acepta emplear "*croscorrelación*", proveniente el término inglés "*cross-correlation*". En esta nota se discute el uso correcto de la técnica de la correlación cruzada en Climatología, enfatizando en su interpretación con un caso concreto, y en el requisito de la estacionariedad de las series, como condición previa y obligatoria a exigir para el cálculo de la croscorrelación, a menos que las series estén cointegradas. Ha sido reconocido, que muchos trabajos que incluyen las técnicas de correlación, correlación cruzada y regresión en diferentes especialidades, carecen de validez, por el descuido o desconocimiento del problema de la regresión espuria, de allí que insistir sobre este problema sea otro objetivo de esta contribución. Granger y Newbold (1974) sugieren que "cuando se modelan regresiones con series de tiempo, si el valor de R^2 es mayor que el del estadístico Durbin-Watson, se debe sospechar la existencia de una relación espuria".

Palabras claves: Correlación cruzada, estacionariedad, croscorrelograma, cointegración, correlación espuria, Taiti[1].

[1] Por cuanto en el DRAE no están registradas Taití ni Tahiti, preferimos, la primera opción al tiempo que coincidimos con Paul Gauguin cuando en su segunda permanencia en la isla en 1897, pintó el cuadro que tituló "Nevermore o Taití"

* Recibido 25.11.2013 Aprobado: 16.01.2014

ABSTRACT

The correlation between two time series, both in simultaneous time and outdated over time is known as crossed correlation and it is commonly accepted as cross-correlation. This paper discusses the proper use of the technique of cross-correlation in Climatology, emphasizing its interpretation in a particular case, and the requirement of stationarity of the series, as a preliminary and required mandatory condition for calculating the cross-correlation, unless series are co-integrated. It has been recognized that many works that include the correlation crossed correlation and regression techniques, in different specialties, lack of validity by the problem carelessness or ignorance of spurious regression, hence insist on this problem is another objective of this contribution. Granger and Newbold (1974) suggest that "when regressions are modeled with time series, if the value is greater than the Durbin-Watson statistic, it must be suspected the existence of a spurious relationships"

Key words: Crossed Correlation, stationarity, croscorrelograma, co-integration, spurious Correlation, Taiti

PRESENTACIÓN

Cuando se desea cuantificar la relación o asociación entre dos series del tiempo o del clima, o entre una de ellas y otra variable de naturaleza no climática, usualmente se recurre a métodos paramétricos como el coeficiente de correlación lineal, como el de *Pearson* o a no paramétricos de Spearman o el de *Mann-Kendall*. En muchas situaciones no hay indicios de correlación entre los eventos expresados por estos coeficientes, puesto que ellos solo expresan la asociación en tiempo simultáneo, pero cuando se les aplica a las series la técnica estadística de la correlación cruzada, es posible que resalten asociaciones muy claras entre las series. Esa correlación cruzada es, entonces, la correlación entre una serie X_t en un tiempo dado, t, con otra serie en un tiempo posterior, Y_{t+k}.

La correlación cruzada entre dos series estacionarias[2], para *lag* (retardo) positivo $r_{xy\,(k)}$ y para lag negativo $r_{yx\,(-k)}$, se obtiene por las fórmulas (1) y (2) y se ejemplifica en el cuadro 1. El coeficiente de croscorrelación en el lag cero, tiene el mismo valor que el coeficiente de correlación lineal de Pearson puesto que las series ocurren simultáneamente, no hay lag (retardo o desfase) entre los datos. En la expresión (1) se lee que la **suma** de los productos de las desviaciones de la variable X_t por las desviaciones de la variable Y_t, desfasada en 1, 2,...(N-1) lags, se divide entre ($N\,S_x\,S_y$) y es lo mismo que decir que la correlación cruzada $r_{xy(k)}$, entre dos series estacionarias para un retardo dado es la **media** de los productos de las series X_t y Y_t normalizadas, desfasadas en 1, 2,...(N-1) lags.

$$r_{XY(K)} = \frac{\Sigma[(X_t - \mu_x)*(Y_{t+k} - \mu_Y)]}{N\,S_x\,S_Y} = \frac{1}{N}\sum[(\frac{X_t - \mu_x}{S_x})*(\frac{Y_{t+k} - \mu_Y}{S_Y})]\ \ en\ \ lag+ \qquad (1)$$

$$r_{YX(-K)} = \frac{\Sigma[(Y_t - \mu_y)*(X_{t+k} - \mu_Y)]}{N\,S_x\,S_Y} = \frac{1}{N}\sum[(\frac{Y_t - \mu_y}{S_y})*(\frac{X_{t+k} - \mu_x}{S_x})]\quad en\ lag- \qquad (2)$$

2 Se dice que una serie temporal es estacionaria cuando su media y su varianza no varíen estadísticamente con el tiempo y la autocovarianza dependa del retardo entre los datos y no del tiempo mismo. La estacionariedad asegura la independencia de los datos, postulado esencial en estadística. Si las series son "no estacionarias" o "integradas", pueden originar correlaciones cruzadas y regresiones espurias. Se exceptúa el caso de series no estacionarias, pero que son "cointegradas", es decir, que entre ellas exista una combinación lineal que sea estacionaria, Orion, Julius (2009); Smith, J.O.(2007); Podobnik Boris y H. Eugene Stanley (2007); Nau (2005); Asteriou (2002).

Donde:

- Xt, Xt+k, valor de un dato en el tiempo, t, y el dato en el tiempo k, en la serie independiente y estacionaria.
- Yt, Yt+k, valor de un dato en el tiempo, t y el dato en el tiempo k en la serie dependiente y estacionaria
- SX y SY, desviación estándar poblacional de las series Xt y Yt, respectivamente
- N, número de pares de datos de las series Xt y Yt.
- μx y μy, medias de las series Xt y Yt, respectivamente
- k, retardo (lag) entre una observación en tiempo t y otra en tiempo posterior o anterior, k: 0, ±1, ± 2,... ±.N-1
- N, SX, SY, μx y μy corresponden a la serie de lag cero y se mantienen constantes en los demás lags

El cuadro 1 resume la fórmula a utilizar según que el retardo sea positivo, sin retardo o con retardo negativo. Los datos que se desplazan en la variable X_t o Y_t determinan el signo positivo o negativo del retardo de los coeficientes. La correlación cruzada, igual que la correlación lineal y la autocorrelación, varía de -1 a +1 para indicar máxima correlación negativa o positiva, respectivamente, pero diferente a la correlación lineal simple, el coeficiente de correlación cruzada entre las series X_t y Y_t difiere si la correlación es entre Y_t y X_t, en lags diferentes de cero.

Significación estadística. Se acepta que un coeficiente de correlación cruzada es **estadísticamente significativo**, (estadísticamente diferente a cero) al 95% de confianza, si se cumple que:

$$r_{xy}(k) > \frac{1{,}96}{\sqrt{(N-k)}} \qquad (4)$$

Donde k es el valor absoluto del lag dado, N, el número de datos, y $1/\sqrt{(N-k)}$ el error estándar de $r_{xy}(k)$. Aplicado al r de lag +1= 0,614, (cuadro 6); como r es mayor que (1,96*0,204) se considera estadísticamente significativo.

Cuadro 1.
Fórmulas de los coeficientes de correlación cruzada, según el signo del retardo, en series con datos mensuales

$r_{XY(K)} = \frac{1}{N}\sum[(\frac{X_t-\mu_x}{S_x})*(\frac{Y_{t+k}-\mu_Y}{S_Y})]$ (1)		Fórmula para Retardos positivos (+)
X_t	Y_{t+k}	Lag 1 (Los datos del mes en X_t se asocian con los de Y_t de 1 mes después)
E →	F	
F →	M	
M →	A	
A →	M	
M →	J	
$r_{XY(K)} = \frac{1}{N}\sum[(\frac{X_t-\mu_x}{S_x})*(\frac{Y_t-\mu_Y}{S_Y})]$ (3)		Fórmula sin Retardo (lag cero)
E →	E	Lag 0
$r_{YX(-K)} = \frac{1}{N}\sum[(\frac{Y_t-\mu_y}{S_y})*(\frac{X_{t+k}-\mu_x}{S_x})]$ (2)		Fórmula para Retardos negativos (-)
X_{t+k}	Y_t	Lag -1 (Los datos del mes en Y_t se asocian con los de X_t de 1 mes después)
E	← E	
F	← F	
M	← M	
A	← A	
M	← M	
J	← J	

Obs: Es los mismo calcular $r_{yx(-k)} = r_{xy(-k)}$ por: X_{t+k} con Y_t; Y_t con X_{t+k} o X_t con Y_{t-k}

LA CROSCORRELACIÓN Y LA ESTACIONARIEDAD DE LAS SERIES

Antes de proceder a calcular la correlación cruzada entre dos series cronológicas, hay que asegurarse de que ambas series sean "estacionarias", es decir, que sus medias y varianzas sean independientes del tiempo; de lo contrario, de ser "no estacionarias" (o "integradas"), los valores de la media, la varianza o ambas -a medida que aumenta el tamaño de la serie- con el pase del tiempo, tienden a cambiar, perdiendo su comportamiento de estabilidad estadística y aunque los coeficientes resultasen altos y significativos, no tendrían validez y las correlaciones podrían ser consideradas espurias, ya que serían esencialmente causadas por la tendencia o por autocorrelación en las series. Es decir, las correlaciones cruzadas de series no estacionarias no constituirían relaciones reales entre las variables, sino relaciones artificiales y carentes de sentido, a menos que X_t y Y_t sean series cointegradas. Véase Box, Gep and Jenkins (1979), Makridakis *et al.* (1983), Gujarati (1999), Smith (2007), Granger (2004), Nau (2005a), Hagg (2005) y Podobnik (2007), entre otros. En cuanto a la normalidad de las series, en la determinación de la correlación cruzada, técnicamente no es necesario que lo sean, pero los residuales sí, Arnaus (2001), Mata, H (s/f). Ahora, si además de estacionaria, los residuales fuesen normales, las series serían estrictamente estacionarias.

Enfatizando: cuando las variables no cumplan con la condición de estacionariedad, pueden ser utilizadas en correlación, croscorrelación y regresión, solamente si están "**cointegradas**", es decir, que siendo variables no estacionarias, entre ellas existe una **combinación lineal** que **sí es estacionaria**, y esa combinación lineal entre ellas es la serie de los residuales de la regresión, tal como lo estableció Granger en 1981 y lo ampliaron Engle y Granger (1987) en lo que se conoce como la teoría de la cointegración. Esta teoría es la base de la metodología que evita que la regresión, la correlación y la croscorrelación sean espurias, al calcularlos, solo si los residuales de la regresión entre las series originales o entre las series transformadas, sean estacionarios.

Con el **autocorrelograma** se identifica preliminarmente si alguna de las series, X_t o Yt, es no estacionaria, y se detecta porque los coeficientes de autocorrelación son altos y significativos en muchos lags y van disminuyendo muy lentamente hasta alcanzar cero. En cambio, si es estacionaria, los coeficientes van decreciendo de manera exponencial y rápidamente alcanzan el valor cero.

Identificada la condición de no estacionariedad en una o ambas series, en la mayoría de los casos, si la serie presenta tendencia determinística, se logra su estacionariedad mediante regresión. Si posee tendencia estocástica, por la

trasformación de **las primeras diferencias**, para obtener una nueva serie de diferencias de valores sucesivos, DX_t o DY_t, y cada nuevo dato expresado por (5), bien para la serie X o para la serie Y:

$$DX_t = (X_t - X_{t-1}) \quad \text{y} \quad DY_t = (Y_t - Y_{t-1}) \quad (5)$$

Donde X_t, Y_t es un valor de la variable en el tiempo t y X_{t-1}, Y_{t-1}, es la variable X o Y en el tiempo anterior, (retardada un lag o unidad temporal empleada).

De no resultar estacionaria la serie de las primeras diferencias, se le aplica a esta serie diferenciada, las primeras diferencias, lo cual es denominado "diferencias de segundo orden" o "primeas diferencias de las primeras diferencias". Las series estacionarias resultantes, sean del primer orden o de segundo orden, serán las series a emplear en la croscorrelación y no las series originales. Es decir, en lugar de X_t y Y_t se correlacionará DX_t con DY_t, puesto que al ser series estacionarias, cumplen con el supuesto de independencia.

Si la serie posee ciclo anual (**estacionalidad**), es no estacionaria por periodicidad, y esta estacionalidad o ciclo anual debe ser removida para convertir la serie en estacionaria, lo cual se logra al convertir los datos originales en anomalías, en relación con las **medias mensuales**.

En cualquiera de los casos, si en el autocorrelograma, las autocorrelaciones de las series X_t y Y_t tienden a no ser significativas, no excederán el valor de las bandas de confianza del 95%, expresado en la fórmula (4), y las series serán estacionarias, o que se logró la estacionariedad de las series.

Aunque, en muchos casos, el uso del autocorrelograma es suficiente para identificar la estacionariedad de una serie cronológica con el propósito del cálculo de la correlación cruzada, numéricamente, la aceptación de estacionariedad de una serie se obtiene comúnmente de manera concluyente mediante la prueba de *Dickey Fuller Aumentada* (ADF) (1979) o la de Phillips-Perron (PP) (1988), ambas en el programa econométrico *Eviews, Quantitative Micro Software* (2002). En estas pruebas, si el valor absoluto del estadístico ADF es mayor que el valor absoluto crítico de MacKinnon a un nivel α dado, se acepta que la serie no posee raíz unitaria y, en consecuencia, es considerada estacionaria. También es empleada la prueba de *Kwiatkosky, Phillips, Schmidt, and Shin* conocida por (KPSS) (1992) en las versiones más recientes del Eviews.

En ausencia de un paquete estadístico, una manera práctica para saber si una serie es **estacionaria** en la media, es subdividirla en varios sub períodos de aproximadamente la misma longitud y se calculan sus medias aritméticas. El proceso

será considerado estacionario si las medias de los sub periodos son aproximadamente iguales (o que entre dichas medias y la media de la serie completa, no hubiese diferencias significativas, lo cual se puede decidir mediante la prueba t de Student para muestras correlacionadas).

Se dijo que la varianza de los residuales debe ser constante con el tiempo, y ello también se aprecia en el gráfico de los residuales de la regresión. Si existiese heteroscedasticidad (cambios en la varianza con el tiempo) es recomendable una transformación logarítmica de la serie original, o bien aplicarle la transformación de *Box-Cox*.

Un caso explicado de correlación cruzada: La presión atmosférica en Taití y Darwin Metodología y datos

Para reafirmar el procedimiento explicado, se calculará la correlación cruzada en diferentes retardos (lags) entre la presión atmosférica mensual reducida al nivel del mar en Papelee, Taití (17°35´S-149°37´W), Polinesia Francesa, y la presión atmosférica mensual reducida al nivel del mar en Darwin (12°26´S–130°52´E), Australia, ambas en el océano Pacifico, durante la presencia de La Niña o fase fría del fenómeno ENSO (El Niño Oscilación del Sur). Las series son expresadas en hectopascales, hPa, y representan adecuadamente los polos opuestos en el concepto del índice de la Oscilación del Sur, IOS, o componente atmosférica del ENSO (Guevara, 2008).

Datos. Los datos empleados de la presión atmosférica al nivel del mar en hectopascales, provienen de CRU (*Climate Research Unit*, 2009).

Pasos a seguir:

- Se construye el cuadro 2 con el orden que deben llevar las columnas de las series originales X_t=T y Y_t=D; las columnas de los **desvíos** de las series mensuales, (dX_t y dY_t), y los productos de esos desvíos, para obtener la correlación cruzada según las fórmulas (1) y (2), para los lags +1 y -1.
- Se considera a la serie Taití como la serie independiente, X_t, precursora, líder, indicativa, en la primera columna del cuadro 2, (a la derecha de las columnas de años y meses) de allí que también se le denomine serie 1, dada su ubicación cercana a la alta presión atmosférica generadora de los vientos alisios. La serie Darwin será considerada la serie dependiente, Y_t, y se coloca en la segunda columna, por ello también es denominada serie 2.

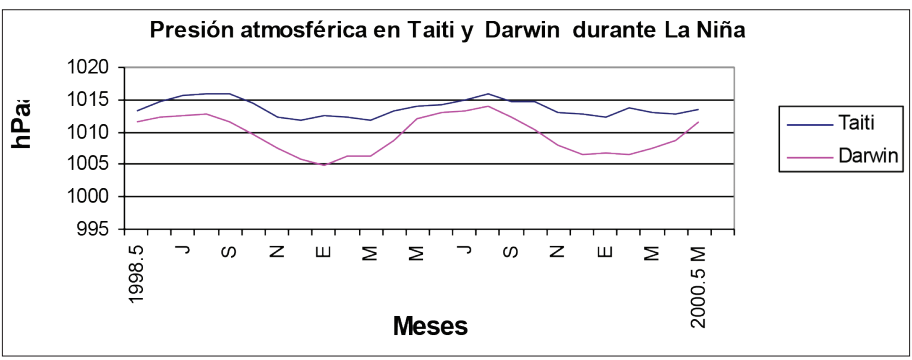
Cuadro 2.
Cálculo de la correlación cruzada entre las presiones atmosféricas reducidas al nivel del mar (hPa) en Taíti y Darwin

		X_t	Y_t	dX_t	dY_t	$(dX_t * dY_{t-1})$	$(dX_{t-1} * dY_t)$
Año	**Mes**	**Taiti**	**Darwin**	**d(Taiti)**	**d(Darwin)**	**Lag+1**	**Lag-1**
1998	5	1013.2	1011.5	**-0.54**	1.89	**-1.41**	1.62
1998	6	1014.6	1012.2	0.86	**2.59**	2.47	5.06
1998	7	1015.7	1012.5	1.96	2.89	6.24	6.52
1998	8	1016	1012.8	2.26	3.19	4.71	6.87
1998	9	1015.9	1011.7	2.16	2.09	0.19	1.37
1998	10	1014.4	1009.7	0.66	0.09	-1.45	-0.14
1998	11	1012.2	1007.4	-1.54	-2.21	5.73	4.08
1998	12	1011.9	1005.9	-1.84	-3.71	8.69	4.25
1999	1	1012.6	1004.9	-1.14	-4.71	3.90	7.28
1999	2	1012.2	1006.2	-1.54	-3.41	4.96	6.29
1999	3	1011.9	1006.4	-1.84	-3.21	1.68	1.75
1999	4	1013.2	1008.7	-0.54	-0.91	-1.35	-0.14
1999	5	1013.9	1012.1	0.16	2.49	0.53	1.38
1999	6	1014.3	1013	0.56	3.39	1.99	3.92
1999	7	1014.9	1013.2	1.16	3.59	4.96	7.74
1999	8	1015.9	1013.9	2.16	4.29	6.01	4.10
1999	9	1014.7	1012.4	0.96	2.79	0.66	2.94
1999	10	1014.8	1010.3	1.06	0.69	-1.70	-0.51
1999	11	1013	1008	-0.74	-1.61	2.24	1.68
1999	12	1012.7	1006.6	-1.04	-3.01	3.04	4.35
2000	1	1012.3	1006.7	-1.44	-2.91	4.49	-0.16
2000	2	1013.8	1006.5	0.06	-3.11	-0.12	2.00
2000	3	1013.1	1007.4	-0.64	-2.21	0.65	2.09
2000	4	1012.8	1008.6	-0.94	-1.01	-1.97	0.15
2000	5	1013.6	1011.7	-0.14	2.09		
					Suma	**55.14**	**74.47**
	Media	**1013.74**	**1009.61**		Coef. CC	**0.614**	**0.829**
	S	**1.29**	**2.79**			**rxy(k=1)**	**rxy(k=-1)**

Fuente de los datos: Cru. Southern Oscillation Index SOI.(2009).
Compare resultados con cuadro de figura 4 calculados mediante el SPSS.

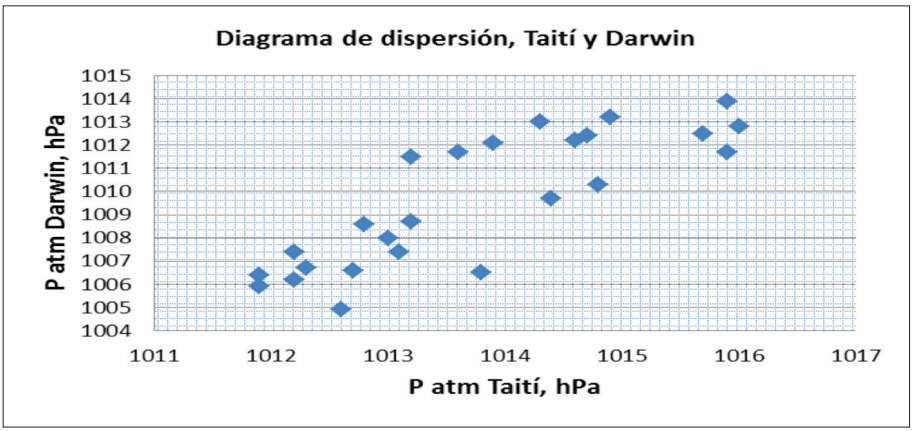
- Se establece claramente lo que se desea encontrar con la croscorrelación. En este caso, como Taití es la variable independiente, precursora, indicativa, se trata de averiguar si esta serie en el tiempo t estará relacionada con la serie Darwin en el tiempo t+k, o bien, responder a la interrogante ¿El incremento de la presión atmosférica superficial en Taití, (expresado por T o X_t) será seguido por incremento en la presión atmosférica superficial en Darwin (D o Y_t)?
- Se debe tener idea del comportamiento las series, individual y comparativamente. Para ello, se grafica la variación de las series con los meses (figura 1a), así como su diagrama de dispersión (figura 1b).
- Se realiza el cálculo de la croscorrelación por las fórmulas (1) y (2) si las dos series son estacionarias, o bien, no estacionarias pero cointegradas de igual orden. Si no cumplen con ninguna de estas condiciones, transformar previamente, las series originales X_t y Y_t en estacionarias, pero las series transformadas deberán ser estacionarias del mismo orden de diferenciación, es decir, si las originales eran de orden uno, las diferenciadas serán de orden cero; si las originales eran de orden 2, las diferenciadas con dos diferenciaciones serán de orden cero.
- Se construye el **croscorrelograma** de las series (figuras 2a y 2b) y se interpreta adecuadamente, recordando que cuando los coeficientes de correlación cruzados están en lags positivos significa que la serie independiente, X_t, lidera a la serie dependiente, Y_t, pero, en lags negativos, la serie Y_t lidera a la serie X_t.

Figura 1a.
Secuencia de las series originales de la presión atmosférica en Taití y Darwin. Indicando inicio de la serie en el mes 5 del año 1998 y finalizando en el mes 5 de 2000



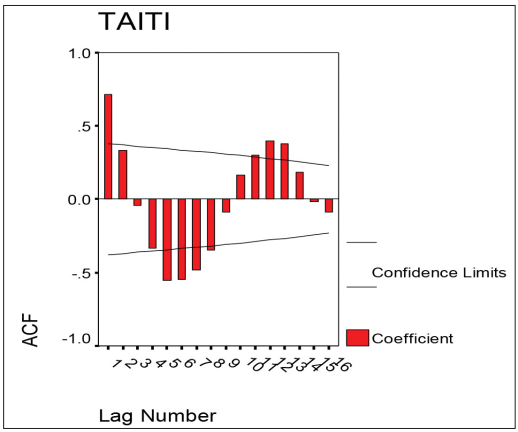
Las secuencias de las series indican curvas cuasi paralelas con fluctuaciones, aunque más acentuadas en Darwin. El diagrama de dispersión (figura 1b) indica **linealidad** entre las series de Taití y Darwin, requisito esencial en regresión y correlación.

Figura 1b.
El diagrama de dispersión de las series de la presión en Taití y Darwin



Los autocorrelogramas de las series de Taiti (figura 2a) y Darwin (figura 2b) muestran la estacionalidad, más acentuada en Darwin, indicativas de la no estacionariedad de ambas series. En los autocorrelogramas (Figura 2a y 2b) los coeficientes de correlación serial son significativos en lags: 1, 2, 4, 5, 6, 7, 8, 10,11, 12 y 13, los cuales sobrepasan los umbrales de significación estadística, representados por las líneas punteadas. La decisión definitiva sobre la estacionariedad o no de las series se realiza mediante dos de las pruebas clásicas de mayor uso: la de ***Dickey Fuller Aumentada,*** (ADF), y la prueba de ***Phillips-Perron*** (PP), con el programa econométrico *Eviews*. La prueba ADF acepta que una serie es estacionaria si el valor absoluto de ADF es mayor que el valor absoluto crítico de MacKinnon al 5%, (u otro valor) y la de la prueba PP, muy similar, pero con diferentes valores críticos de comparación.

Figura2a.
Autocorrelograma de la serie de presión atmosférica en Taití indicando estacionalidad

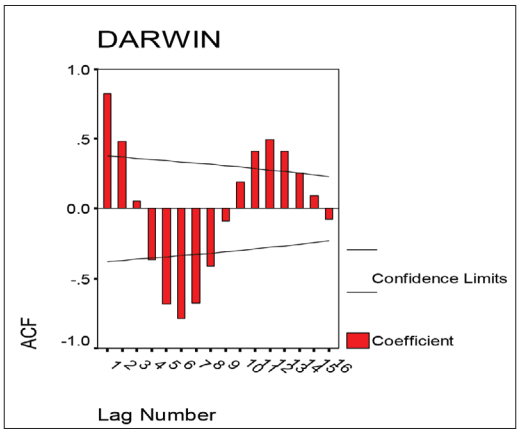


Cuadro 3.
T, autocorrelaciones

Lag	Ac.	Error S	Box-Ljung	Prob.
1	.716	.189	14.425	.000
2	.335	.185	17.716	.000
3	-.044	.181	17.776	.000
4	-.331	.176	21.306	.000
5	-.556	.172	31.723	.000
6	-.548	.168	42.391	.000
7	-.482	.163	51.098	.000
8	-.348	.159	55.914	.000
9	-.089	.154	56.245	.000
10	.165	.149	57.477	.000
11	.303	.144	61.890	.000
12	.399	.139	70.145	.000
13	.382	.133	78.334	.000
14	.187	.128	80.473	.000
15	-.018	.122	80.495	.000

Los estadísticos de Box Ljung en cuadro 3, indican que si su probabilidad es menor que 0,05, lo coeficientes de autocorrelación son significativamente diferentes de cero.

Figura 2b.
Autocorrelograma de la serie de presión atmosférica en Darwin indicando estacionalidad



Cuadro 4.
D, Autocorrelación.

Lag	Corr.	Err.	Box-Ljung	Prob.
1	.825	.189	19.124	.000
2	.485	.185	26.038	.000
3	.057	.181	26.136	.000
4	-.369	.176	30.508	.000
5	-.682	.172	46.201	.000
6	-.786	.168	68.132	.000
7	-.678	.163	85.385	.000
8	-.413	.159	92.143	.000
9	-.089	.154	92.476	.000
10	.192	.149	94.141	.000
11	.414	.144	102.394	.000
12	.493	.139	115.014	.000
13	.408	.133	124.400	.000
14	.256	.128	128.415	.000
15	.095	.122	129.018	.000

Prueba ADF, modelo con constante: -2,58, y valor crítico al 5% -2,99, con 1 variable dependiente retardada, clasifica a Taití como no estacionaria.

Prueba ADF, modelo con constante: -4,29, valor crítico al 5% -3,00 con 2 variables dependientes retardadas, clasifica a Darwin como estacionaria.

Sin embargo, dado que por la prueba de estacionariedad de ***Phillips-Perron*** **(PP)**, en el mismo Eviews no se cumple con la condición de estacionariedad en ninguna de las localidades, se aceptó la decisión de no estacionariedad por esta prueba que, además, concuerda con los autocorrelogramas.

Aceptar la no estacionariedad de ambas series significa que no podrían utilizarse para obtener la correlación cruzada ni tampoco establecer regresión entre ellas, sin el peligro de obtener resultados espurios, es decir, resultados altos, pero sin sentido por no tener significación estadística válida. ¿Qué hacer, entonces? Dos opciones posibles: a) Aunque las series son no estacionarias (integradas), demostrar que entre ellas existe una relación lineal que sí es estacionaria, o sea, demostrar que son series **cointegradas**. Esto se logra al establecer una regresión con las series originales de Taití y Darwin, y si los residuales de la regresión (que es la combinación lineal entre las series) son estacionarios, se aceptará la cointegración de las dos series y se aceptará su empleo en la correlación cruzada. b) La segunda opción es aplicar una trasformación matemática a las series, y si estas series trasformadas son estacionarias, podrán ser empleadas en la croscorrelación.

Comprobación de la cointegración de las series:

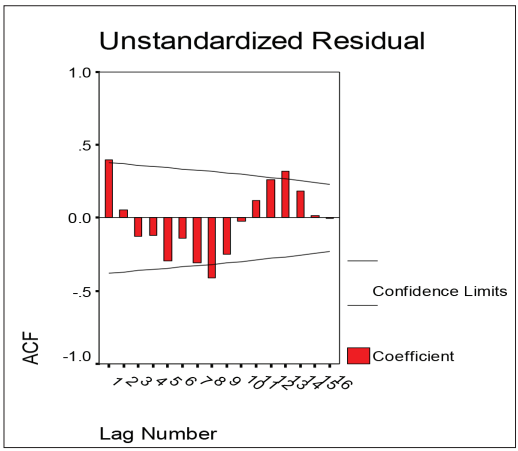
Se obtiene la siguiente regresión lineal entre la presión atmosférica en Taití ($P_{Taití}$) y Darwin (P_{Darwin}), ambas no estacionarias

$$P_{Darwin} = -828{,}65 + 1{,}81\ P_{Taití} \qquad (6)$$

(N: 25; R^2: 0,704; Se:1.57; Error est: 0.24 ; t: 7.4 Pv: 0.00)

En la figura 3 el autocorrelograma de los residuales (*Unstandardized Residual*) de la regresión y los valores de los coeficientes de autocorrelación (AC) correspondientes en el cuadro 5.

Figura 3.
Autocorrelograma de los residuales indicando su estacionariedad



Cuadro 5.
AC de residuales

Lag	Ac.	SErr.	Box-Ljung	Prob.
1	.401	.189	4.525	.033
2	.055	.185	4.612	.100
3	-.129	.181	5.122	.163
4	-.121	.176	5.592	.232
5	-.294	.172	8.502	.131
6	-.138	.168	9.177	.164
7	-.306	.163	12.694	.080
8	-.411	.159	19.390	.013
9	-.248	.154	21.988	.009
10	-.022	.149	22.009	.015
11	.120	.144	22.702	.019
12	.262	.139	26.263	.010
13	.324	.133	32.153	.002
14	.182	.128	34.195	.002
15	.014	.122	34.208	.003
16	-.001	.115	34.208	.005

Prueba **ADF de estacionariedad de los residuales,** mediante Eviews:

Seleccionando "*level*" "*None*" (sin tendencia, sin intercepto) con una variable dependiente retardada. **ADF= -2,71; valor crítico al 5% -1,95**

Prueba PP de estacionariedad de los residuales mediante Eviews:

Seleccionando "*level*" "*None*" (sin tendencia, sin intercepto) con una variable independientes retardada: **PP= -3,13; valor crítico al 5% -1,95**

Decisión: Como el valor absoluto de ADF |-2,71|; y el valor absoluto de PP |-3,13| son mayores que el valor absoluto del valor crítico,|-1,95|, se acepta que los residuales de la regresión son estacionarios, lo que significa que las series están cointegradas y poseen una relación estable en el largo plazo; en consecuencia, aunque las series mensuales de presión atmosférica en Taití y en Darwin utilizadas, resultaron no estacionarias pero cointegradas, **pueden ser utilizadas para calcular la correlación cruzada entre ellas**, sin necesidad de remover la estacionalidad anual que causa la no estacionariedad en ambas series.

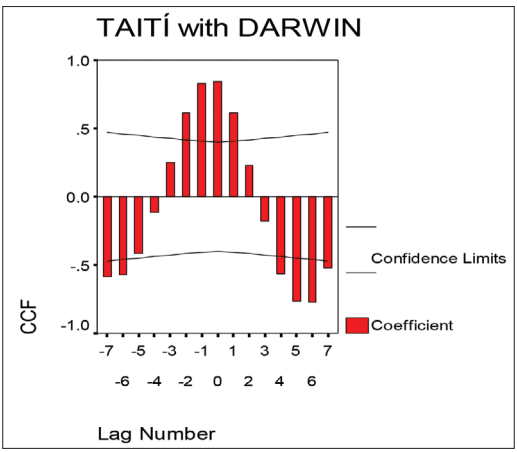
Aplicando las fórmulas (1) y (2) y siguiendo el procedimiento indicado en el cuadro 2, se calculan los coeficientes de correlación cruzada con retardos desde el lag -7 hasta el lag +7 (cuadro 6 y el croscorrelograma respectivo, figura 4), según el programa SPSS. Estos resultados coinciden con los obtenidos directamente online por el calculador de Wessa (2009).

Interpretación del croscorrelograma

La figura 4 es el correlograma de la correlación cruzada **y** representa todas las correlaciones cruzadas del cuadro 6, para visualizar la influencia del desfase temporal en las asociaciones entre las dos variables. El lag cero está en el centro del croscorrelograma. Hacia la derecha, los lags o retardos positivos 1, 2, 3, 4, 5, 6, 7, en meses, que indican que la variable independiente, X_t, lidera el proceso, y hacia la izquierda del lag 0, los coeficientes en retardos negativos, que indican que la variable dependiente, Y_t, lidera el proceso.

Se considera que un coeficiente está en **retardo positivo** cuando los datos de la variable dependiente, Y_t, ocurren en una o más unidades temporales después que los datos de la variable independiente, X_t. Por ejemplo, si la presión atmosférica en X_t es en enero, la de Y_{t+1} será en febrero, un mes más tarde, en relación con la de X_t. (atrasada, retardada, después). Se considera que un coeficiente está en **retardo negativo** cuando los datos de la variable dependiente, Y_t, ocurren en una o más **unidades temporales**

Figura 4.
Croscorrelograma de las series de Taití y Darwin.



Cuadro 6.
Croscorrelaciones

Lag	Corr.	Er E
-7	-0,582	0,236
-6	-0.574	0.229
-5	-0.412	0.224
-4	-0.113	0.218
-3	0.252	0.213
-2	0.611	0.209
-1	0.829	0.204
0	0.840	0.200
1	0.614	0.204
2	0.229	0.209
3	-0.177	0.213
4	-0.562	0.218
5	-0.763	0.224
6	-0.773	0.229
7	-0.522	0.236

En la ordenada, escala de los coeficientes, y en la abscisa, los lags positivos y negativos. Las líneas horizontales punteadas, los límites de confianza del 95%.

A la derecha, el cuadro 6, con los coeficientes de correlación cruzada calculados por el SPSS desde lag -7 a +7 y sus respectivos errores estándar.

antes que los datos de la variable independiente, X_t. Por ejemplo, si la presión atmosférica en X_t es en enero, la de Y_{t+1} será en diciembre, un mes antes en relación con la de X_t. (Antes, adelantada, retardada negativamente).

La variable independiente, X_t, es la presión atmosférica reducida al nivel del mar, en la localidad de Taití, océano Pacífico. La variable dependiente, Y_t, es la presión atmosférica reducida al nivel del mar, en la localidad de Darwin, Australia.

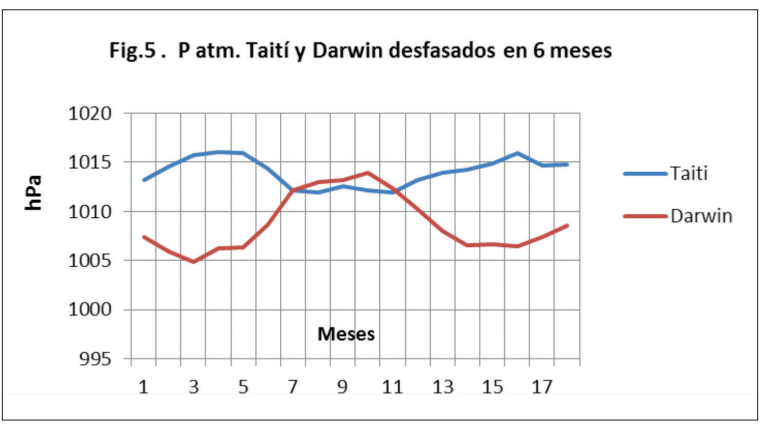
Las series utilizadas corresponden a un periodo de 25 meses durante el cual estuvo presente la Niña, o fase fría del fenómeno El Niño-Oscilación del Sur (ENSO) y el número de coeficientes calculados fue de siete aunque, usualmente, para obtener el croscorrelograma se calcula ¼ del número de datos. En el croscorrelograma de

la figura 4, el coeficiente de mayor valor ocurre en el lag 0 con +0,84, correlación alta y estadísticamente significativa, al sobrepasar la línea del límite de confianza del 95% (fórmula 4). Este valor significa que la presión atmosférica reducida al nivel del mar, aumenta o disminuye **simultáneamente** en Taití y Darwin. Es decir, durante la Niña, cuando la presión atmosférica aumenta en Taití, también lo hace en Darwin, y cuando disminuye en Taití, disminuye en Darwin.

Con un desfase o retardo positivo de un mes entre las series (lag positivo 1), el coeficiente es +0,614, estadísticamente significativo, que al estar en lag positivo, la variable que lidera el proceso es la X_t, Taití, interpretándose que el incremento o descenso bárico en Darwin, se inicia un mes después que el aumento o descenso ocurrido en Taití, es decir, el aumento o el descenso de presión en Taití **precede** en un mes al aumento o descenso de presión en Darwin, durante la presencia de La Niña en el océano Pacífico.

En el croscorrelograma (figura 4) se ve que con desfases de cuatro, cinco, seis y siete meses desde el inicio de la serie en Taití (lags positivos), el proceso de la presión atmosférica entre las dos localidades, durante la Niña tienen otro comportamiento, ahora existe una relación inversa: al aumentar o disminuir la presión en Taití, a los 4, 5, 6 y 7 meses después, respectivamente, la presión atmosférica disminuye o aumenta en Darwin. Ese comportamiento inverso en desfase de 6 meses, por ejemplo, es expresado por la correlación cruzada alta y negativa de -0,773 (cuadro 6) y por la figura 5.

Figura 5.
Comportamiento de la presión atmosférica en Taití y Darwin durante La Niña y con desfase de 6 meses



En lags negativos, (cuadro 6 y figura 4) se observa una repetición de lo que ocurre en los lags positivos, pero con interpretación opuesta ya que la variable que lidera el proceso es la variable Y_t, Darwin. En los desfases negativos de 1 y 2 meses, los coeficientes son positivos y estadísticamente significativos, expresando que cuando la presión atmosférica en Darwin durante La Niña, aumenta, 1 y 2 meses después, se incrementa en Taití.

Con desfase de 6 y 7 meses, desde el inicio de la serie en Darwin, cuando la presión aumenta o disminuye, 6 y 7 meses después, disminuye o aumenta en Taití. Las correlaciones cruzadas en estos desfases son -0,57 y -0,58, respectivamente.

Enfatizando sobre el cálculo de los coeficientes de correlación cruzada

Los programas estadísticos y econométricos resuelven los cálculos de manera rápida y efectiva, pero parte de nuestro interés es enfatizar la realización de los cálculos paso a paso, pensando en los que se inician en el tema de la correlación cruzada como una herramienta imprescindible en las ciencias ambientales, para que los cálculos sean correctamente entendidos y los resultados aplicados adecuadamente. Así, si un coeficiente empieza con la **variable independiente**, X_t, en lag positivo, se escribe $r_{xy(k)}$ y si por ejemplo, la croscorrelación entre la lluvia diaria y el caudal de un río fuese +0,52 en lag de dos días, ($r_{xy(2)} = 0,52$) significa que el caudal del río aumentará dos días **después** de ocurrida la lluvia. O bien, cuando el intenso El Niño de 1988 incrementó la temperatura media global hasta 6 meses, en términos de correlación cruzada, se dirá que el incremento de la temperatura global siguió a El Niño durante seis meses.

Regresemos, entonces, al cuadro 1. Allí están las fórmulas para obtener los coeficientes en los lags positivos, negativos y sin retardo, con flechas señalando la dirección de la asociación entre los datos mensuales de las variables. Con lapsos diferentes la explicación es similar. En los lags positivos, los datos de la variable **independiente, X_t** permanecen fijos, para asociar su primer dato, en enero, X_1, con el dato del mes siguiente de la variable dependiente, en febrero, $Y_{t+1} = Y_2$, en la columna derecha. Luego, el dato de febrero en X_t, X_{t+2}, se combina con el de marzo en Y_t, Y_{t+3}, y así sucesivamente, hasta que el penúltimo de X_t, se combine con el último de Y_t.

Si se recalculase manualmente el coeficiente de correlación cruzada en desfase positivo de 6 meses (curvas representadas en figura 5) se constatará que **el número de datos y la desviación estándar de las series son los mismos para los coeficientes en el lag 1, -1 y para los cálculos de todos los coeficientes**, y esta es una de las razones por las cuales las series a usar en correlación cruzada deben ser estacionarias y con suficiente número de datos.

En el cálculo del coeficiente en lag -1, permanece fija la variable Y_t, y la variable independiente, X_t, es la que se desplaza o desfasa. El dato de enero de $Y_{t\,=}\ Y_1$ se combina con el dato de febrero de la variable $X_{t,=}X_2$. El de febrero de Y_2 con el de marzo de X_3, hasta que el penúltimo de Y_t se combine con el último de X_t. El proceso continua hasta los lag positivos y negativos 2, 3, 4, etc, hasta calcular los coeficientes necesarios, generalmente, hasta N/4 coeficientes.

En el cuadro 2, se identifican las columnas de las series originales y sus desvíos. El desvío de la serie X_t, (dx_t) se multiplica por el desvío de la serie Y_t (dy_t) para formar el primer dato de la columna de los productos de los desvíos. Estos productos se suman y esta suma se divide entre el producto: ($N\,S_x\,S_y$). Remplazando en las fórmulas (1 y 2) con los datos requeridos, se obtiene el coeficiente de correlación cruzada del primer lag positivo (+0,614) y del primer lag negativo (+0,829) entre las series X_t y la serie Y_t, valores que coinciden con los del cuadro 6 obtenido por Wessa (2009) y el SPSS, con indicaciones de sus significaciones estadísticas.

$$r_{XY(K=1)} = \frac{\Sigma[(X_t - \mu_x)*(T_{t+k} - \mu_Y)]}{NS_X\,S_Y} = \frac{55{,}14}{25*1{,}29*2{,}79} = +\,0{,}614 \qquad (7)$$

$$r_{XY(K=-1)} = \frac{\Sigma[(Y_t - \mu_y)*(X_{t+k} - \mu_Y)]}{N\ S_X\,S_Y} = \frac{74{,}47}{25*1{,}29*2{,}79} = +\,0{,}829 \qquad (8)$$

José Manuel Guevara Díaz

CONCLUSIONES Y RECOMENDACIONES

- En el cálculo del coeficiente de correlación cruzada, denominado en inglés "*cross-correlation*" y castellanizado en la practica como crosscorrelación y croscorrelación, siempre es conveniente utilizar como primera variable, a la variable X_t, a la que se considere o sospeche que sea la variable independiente, precursora, líder o causal, y como variable Y_t o segunda variable, la que se considere como variable dependiente. Si se invierte este orden, los coeficientes mantienen sus magnitudes, signos y lags, pero **cambiarán los signos de los lags**, y como consecuencia, el coeficiente máximo se ubicará en el lag de signo opuesto.
- Si las series fuesen anomalías de los datos originales (en relación con las medias de sus series), los valores de los coeficientes no cambiaran, ya que las anomalías de las anomalías, y la desviación estándar de las anomalías, son las mismas que las de las series originales.

- En lag cero, el valor del coeficiente es el mismo que el coeficiente de correlación de Pearson y significa que los eventos ocurren de **manera simultánea,** puesto que no hay separación de tiempo (*lag*) entre ellos.
- Para aplicar la correlación cruzada es necesario que las series sean **estacionarias, o si no lo fuesen, que sean cointegradas**. Por ello se recomienda, previamente, establecer una regresión lineal entre las series, y si sus residuales son estacionarios, las series originales estarán cointegradas y con ellas se puede calcular la croscorrelación, aunque sean no estacionarias. Mediante el autocorrelograma y por las pruebas de estacionariedad de *Dickey Fuller Aumentada* (ADF) y la de ***Phillips-Perron*** **(PP)**, se conoce la estacionariedad de los residuales y de las series originales. Makridakis *et al.* (1983) exigen que la serie independiente sea "**ruido blanco",** aleatoria, para lo cual emplean la transformación del "preblanqueo". Pero, si se sigue el criterio de H. Mata (s/f) de que ruido blanco es igual a estacionario, habría coincidencia con Makridakis *et al* (1983).
- Se debe tener seguridad en la escogencia de las variables: la variable líder o precursora, se identifica en el croscorrelograma por el **signo del *lag* que** es **positivo y significa que ella lidera o conduce a la serie Y_t. Si el signo del lag es negativo, la variable líder o precursora, es la serie Y_t, y significa que ella lidera o conduce a la serie X_t.**
- Es conveniente que los programas estadísticos se empleen después de entendido el proceso de la correlación cruzada de manera manual. Para el cálculo en la primera columna debe colocarse la serie X_t (1, independiente, precursora, causal o la que se sospeche que lo sea) y a la derecha, la serie de la variable dependiente, Y_t o 2.
- En muchos casos solo es posible la correlación cruzada X_t y Y_t. Por ejemplo, si El Niño y la malaria en un área es +0,60 en lag de 6 meses, significa que "la malaria ocurrió 6 meses después de iniciado el fenómeno El Niño" o "la malaria sigue a El Niño ocurrido 6 meses antes". Si la correlación fuese 0,60 en lag -6, el coeficiente se calculó con la malaria como variable independiente, sin serlo, y sería un contrasentido decir que "El Niño se inició 6 meses después que ocurrió la malaria". O que ¡la malaria lidera El Niño!
- Un gran número de trabajos que emplean correlación cruzada no cumplen con las condiciones exigidas y entre las cuales se enumeran las siguientes: no emplean series estacionarias o si lo son, no explican el procedimiento empleado de estacionarización; no titulan adecuadamente el croscorrelograma; las variables que se utilizan no se especifican claramente, no se sabe si son series cronológicas originales o transformadas; los croscorrelogramas carecen de líneas de significación; la interpretación del croscorrelograma generalmente

es incompleto o poco explicativo y el caso extremo, solo indican el valor de la correlación con su lag, lo demás queda a la imaginación.

- Se recomienda investigar la afirmación de Nau (2005b) de solo aceptar valederos los coeficientes de croscorrelación de los lags 0, 1 y 2, ya que por la condición de "ruido" que contiene la croscorrelación, los coeficientes más allá del tercer lags son puramente accidentales, a excepción de cuando la influencia estacional es importante. Igualmente "ruidosos" son los coeficientes de autocorrelación, considerando probablemente, no ciertos los coeficientes en los lags 5 y 7, en datos mensuales.

REFERENCIAS BIBLIOGRAFÍCAS

ARNAUS, GRASS, J (Ed). (2001). *Diseño de series temporales: Técnicas de Análisis*. Documento en Línea. Ediciones Universitat Barcelona. Barcelona, España. Disponible en: http://books.google.com [Consultado: 15-1-2009].

ASTERIOU, DIMITRIOS (2002). *Notas sobre Análisis de Series de Tiempo: Estacionariedad, Integración y Cointegración*. Documento en Línea. Traducida por H Mata. Disponible en: http://www.personal.rdg.ac.uk/~less00da/lecture3.htm [Consultado: 11-4-2010].

BOX, GEP AND JENKINS, GM. (1979). *Time series analysis: forecasting and control*. San Francisco. Holden-Day, 1976.

CLIMATE RESEARCH UNIT (CRU) (2009). *Data: Southern Oscillation Index (SOI)*, Documento en Línea. UK. http://www.cru.uea.ac.uk/ftpdata/soi.dat [Consultado: 10-5-2009].

EVIEWS v. 2.0. ECONOMETRIC (1997). The Eviews Help. Documento en Línea. Dipsonible en: http://onlinelibrary.wiley.com/doi/10.1111/1467-6419.00026/abstract

ENGLE, R. y GRANGER, C.W. (1987). Cointegration and Error Correction: Representation. Estimation and Testing. *Econométrica*. Vol. 55, pp 251-276.

GRANGER, C. W. J. AND P, NEWBOLD (1974). Spurious Regressions in Econometrics. *Journal of Econometric*. Vol. 2, N° 2, pp 111-120.

GRANGER, C.W.J. (1981). Some Properties of Time Series Data and Their Use in Econometric Model Specification. *Journal of Econometrics*. Vol. 16, N° 1. pp 121-130.

GRANGER, C.W.J (1986). Developments in the Study of Cointegrated Economic Variables *Oxford Bulletin of Economics and Statistics*. Vol. 48. N° 3, pp. 213-227.

GRANGER, CLIVER W. J (2004). Análisis de series temporales, cointegración y aplicaciones. *Revista Asturiana de Economía* - RAE. Nº 30.

GUEVARA DÍAZ, JOSÉ M. (2008). El ABC de los índices usados en la identificación y definición cuantitativa de El Niño - Oscilación del Sur (ENSO). *Terra. Nueva Etapa.* Vol, XXIV, No. 35, pp 85-140.

GUJARATI, DAMODAR (1999). *Econometría*. 3ª ed. Mc Graw Hill, Bogotá.

MAKRIDAKIS, SPIROS; STEVEN WHEELWRIGHT AND VICTOR MCGEE. (1983). *Forecating. Methods and application. Second edition*. John Wiley and Sons, New York.

MATA, H. L. (s/f). *Nociones elementales de cointegración*. Enfoque de Engle-Granger. Documento en Línea. Disponible en: http://webdelprofesor.ula.ve.hmata. [Consultado: 12-5-2008].

NAU, ROBERT. (2005a). *Stationarity and differencing.* Decisión 411 Documento en Línea. Disponible en: http://www.duke.edu/~rnau/411diff.htm [Consultado: 10-3-2009].

NAU, ROBERT. (2005b). *Fitting time series regression models.* Decisión 411. Documento en Línea. Disponible en: http://www.duke.edu/~rnau/timereg.html. [Consultado: 1-4-2009].

ORION, JULIUS. (2009). Cross-Correlation en DSP Relat.com. Documento en Línea. Disponible en: http://www.dsprelated.com/dspbooks/mdft/Cross_Correlation.html. [Consultado: 21-4-2010].

PODOBNIK, BORIS y EUGENE STANLEY (2007). *Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Non-stationary Time Series.* Documento en Línea. Disponible en: http://arxiv.org/PS_cache/arxiv/pdf/0709/0709.0281v1.pdf [Consultado: 1-3-2011].

SMITH, J.O. (2007). *Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications.* Documento en Línea. Second Edition. Disponible en Línea: http://ccrma.stanford.edu/~jos/mdft/. Cross_Correlation.html [Consultado: 10-1-2008].

WESSA, P. (2009). *Free Statistics Software.* Documento en Línea. Office for Research Development and Education. Versión 1.1.23-r3. Disponible en Línea: http://www.wessa.net/ . [Consultado: 15-5-2009].

José Manuel Guevara Díaz. Egresado de la Universidad Central de Venezuela como Licenciado en Geografía, obtiene maestría en la Universidad de Boston y posteriormente, doctorado en la UCV. Profesor Titular. Durante su labor educativa universitaria de pregrado ha dictado los cursos de Meteorología, Climatología, Geografía Física, Geografía de Venezuela, Geografía Regional y trabajo de Campo. A nivel de postgrado, Climatología Urbana y Problemas Climáticos de Venezuela. Su obra escrita es diversa con artículos publicados en revistas nacionales y extranjeras y en libros como: *Meteorología*; *La Geografía Regional, la Región y la Regionalización*; *Métodos de estimación y ajuste de datos climáticos*, *Geografía de las regiones Central y Capital de Venezuela* y su más reciente, *Historia de la Escuela de Geografía de la UCV.*

Correo Electrónico: