Documento de Trabajo 95–05
Serie de Estadística y Econometría 02
Mayo de 1995

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

PREDICCIÓN CON DATOS FALTANTES: APLICACIÓN A UN CASO REAL

Pedro Delicado y Ana Justel [1]

Resumen

En este artículo se realiza un estudio comparativo de modelos lineales y modelos mixtos, mezcla de un componente lineal y un componente no lineal en la parte estacional, para predecir la altura significativa de ola. Los datos proceden de una boya situada en el mar Cantábrico que registra la altura de ola cada tres horas. El interés central es obtener predicciones a corto plazo, dos días, que permitan advertir del estado de mar a los puertos. El principal problema que presenta esta serie para su modelización es el alto porcentaje de datos faltantes. Se completa la serie con un interpolador lineal óptimo, en el sentido de minimizar el error cuadrático medio.

Palabras Clave:
Altura significativa de olas. Error cuadrático medio. Interpolación lineal. Modelos ARIMA. Suavizado no paramétrico.

[1] Departamento de Estadística y Econometría, Universidad Carlos III de Madrid.

1. INTRODUCCIÓN

El estudio de fenómenos físicos permite disponer de grandes bases de datos que contienen mediciones de magnitudes relevantes. En muchos casos estas medidas se registran en intervalos regulares de tiempo. Se construyen así series temporales que presentan características particulares debidas al fenómeno físico a que hacen referencia y al instrumental empleado en su cuantificación.

En este trabajo se estudia la serie de alturas de ola registradas en una boya situada en el mar Cantábrico. El objetivo principal es desarrollar una metodología que pueda ser aplicada a otras series de oleaje con características similares a ésta. El interés se centra en modelizar estas series para realizar predicciones a corto plazo, uno o dos días. Este horizonte de predicción es importante para las autoridades portuarias, ya que les permiten alertar a la flota de posibles variaciones del estado de la mar. Además, es deseable que las predicciones se realicen con rapidez para que su conocimiento sea útil. Por esta razón se ha optado por la modelización univariate de la serie de alturas, pese a que las predicciones podrían mejorar sensiblemente si se utilizase un modelo estructural que incluyese información de temperaturas, vientos, etc. Como características particulares de la serie cabe mencionar su gran tamaño (14608 datos), su periodicidad poco habitual ($s = 2920$ datos, un año), la posible existencia de estacionalidad diaria y el gran número de datos que no se han observado debido a las averías de los aparatos de medida y a fallos en la transmisión de datos.

La línea de trabajo seguida en el análisis de la serie de alturas de ola se puede dividir en tres fases. En primer lugar se procede a la interpolación de datos faltantes. Una vez reconstruida la serie se procede a la identificación y estimación de los modelos. Por último se selecciona el modelo que dé lugar a mejores predicciones a corto plazo fuera de la muestra. Estas predicciones se comparan con los datos del último mes obseravado que no se utiliza en la fase de estimación.

La ausencia de un gran número de datos en la serie (aproximadamente un 13%) y su distribución en el tiempo (en general los datos faltantes están aislados y los intervalos entre ellos son relativamente cortos, pero también hay algunos períodos largos en los que

no se registró ningún dato) hacen que tanto la identificación de modelos como su eventual estimación se vea dificultada enormemente. Si el modelo seleccionado para realizar las predicciones incluye información de los datos registrados en instantes de años pasados, la falta de un dato impedirá predecir el valor de la serie en los mismos instantes de años posteriores. La interpolación inicial de los datos faltantes con la media de todas las observaciones registradas el mismo día en otros años permite identificar un modelo. Con este modelo como base se aplica la técnica de interpolación óptima de datos faltantes propuesta por Maravall y Peña (1992).

La modelización de la serie interpolada se realiza en dos direcciones. La primera consiste en aplicar la metodología Box-Jenkins (Box y Jenkins, 1976) para la estimación de modelos lineales ARIMA estacionales. Estos modelos han sido ampliamente tratados en la literatura de series temporales en áreas como la economía, demografía o la física. La segunda se basa en una modelización mixta, con una parte no lineal para recoger la estacionalidad y otra lineal que recoge la estructura de dependencia regular mediante un modelo ARIMA.

El problema del porcentaje alto de datos faltantes se tratará en la sección 2. En la sección 3 se proponen dos tipos de modelos, uno lineal y otro mixto y en la sección 4 se comparan las predicciones a corto plazo de ambos modelos. Finalmente, en la sección 5 se incluyen algunos comentarios finales.

1.1. Altura significativa de ola

La altura de una ola se define como la distancia entre el valor mínimo de la ola, *valle*, y el valor máximo, *cresta*. La medición del oleaje se suele realizar con boyas que disponen de dispositivos para medir las aceleraciones que las olas comunican a la boya, esta medición se realiza durante un periodo de 30 minutos aproximadamente y en intervalos cortos de tiempo. El sensor de a bordo integra dos veces la serie de aceleraciones y se obtiene la serie de elevaciones de ola o estado de mar. La elevación de la superficie del mar a corto plazo es un proceso estocástico estacionario en media, ya que se supone que el nivel del mar es constante, y en varianza.

Durante muchos años la información del oleaje se obtenía a partir de registros visuales de barcos en ruta. La percepción humana tiende a sobrevalorar la altura de las olas y, por tanto, para poder contrastar la información histórica con los datos actuales es necesario definir un parámetro relativo al estado de mar que pueda ser comparable a la altura de ola registrada visualmente. El parámetro comúnmente utilizado es el valor medio del tercio de alturas más altas del registro. Este valor se conoce con el nombre de ***altura significativa de ola*** y se denota por h.

A partir de la serie de elevaciones de la ola η_t se estima la altura significativa de ola h mediante la densidad espectral $f(\nu)$, ya que

$$Var(\eta_t) = \sigma^2 = \int f(\nu)d\nu,$$

y la altura significativa de ola se puede aproximar por 4σ,

$$h \approx 4\left(\int f(\nu)d\nu\right)^{1/2}.$$

En la práctica se realiza un control de la calidad del registro de elevaciones para comprobar que es realmente estacionario, posteriormente se calcula la densidad espectral $f(\nu)$ mediante la transformada rápida de Fourier (FFT). Finalmente, la altura significativa de ola h se obtiene integrando numéricamente $f(\nu)$. Para más detalles sobre la construcción de la serie de oleaje se pueden consultar los libros de Goda (1985) y Sorensen (1993).

Los datos de la serie h_t que se estudia en este artículo proceden de las mediciones registradas en una boya situada cerca de Gijón. Cada 0.5 segundos se mide la aceleración en un total de 5120 instantes, lo que supone un periodo de observación de aproximadamente 42 minutos. Estos registros se realizan cada tres horas dando lugar a la serie de altura significativa de ola mediante el proceso descrito anteriormente. Los datos disponibles son las alturas significativas de ola cada tres horas en el periodo que abarca desde el 1 de enero de 1986 al 31 de enero de 1991. En este periodo se contabilizan un total de 1871 observaciones faltantes, lo que supone un 13 por ciento del total de los datos, que deberían ser 14608. El gráfico de la serie se muestra en la figura **1**, donde los datos faltantes se reflejan en la línea inferior. Los box-plots de toda la serie y la serie por años reflejan la asimetría de los datos (ver figura **2**a y **2**b). Los box-plots de la figura **2**c corresponden a

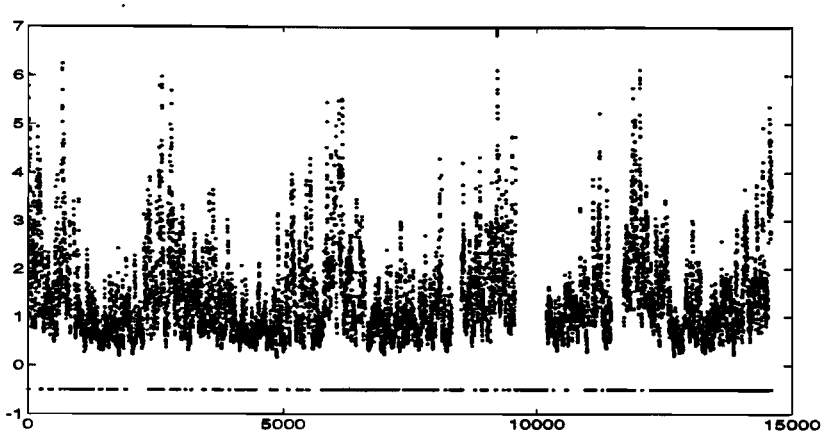


Figura 1: Serie de altura significativa de ola.

los datos registrados en cada mes del año. La figura **2**d representa la serie temporal de los box-plots que corresponden a los datos agregados mensualmente. La serie presenta mayor variabilidad en los meses de invierno. También se han construido los box-plots para los datos correspondientes a cada hora del día pero no se muestran ya que no presentan características diferentes a las que se observan en el box-plot de la serie completa. Debido al comportamiento heterocedástico y asimétrico que se observa en el gráfico de la serie optamos por transformar los datos. A partir de ahora la serie con la que trabajamos es $z_t = \log h_t$, el logaritmo de la serie de alturas significativas de ola que se muestra en la figura **3**.

2. INTERPOLACIÓN DE DATOS FALTANTES

La ausencia de observaciones en series temporales dificulta la estimación de la función de autocorrelación, reduciendo el número de términos con los que se calcula el estimador de cada coeficiente de correlación. La función de autocorrelación (ACF), junto con la función de autocorrelación parcial (PACF), son las principales herramientas que se utilizan en la fase de identificación de modelos ARIMA.

A pesar de la ausencia de 1871 datos en la serie z_t se puede construir su correlograma gracias a que se dispone de muchas observaciones. Sin embargo, debido al comportamiento no estacionario de la serie es necesario diferenciar z_t con una diferencia regular y otra de

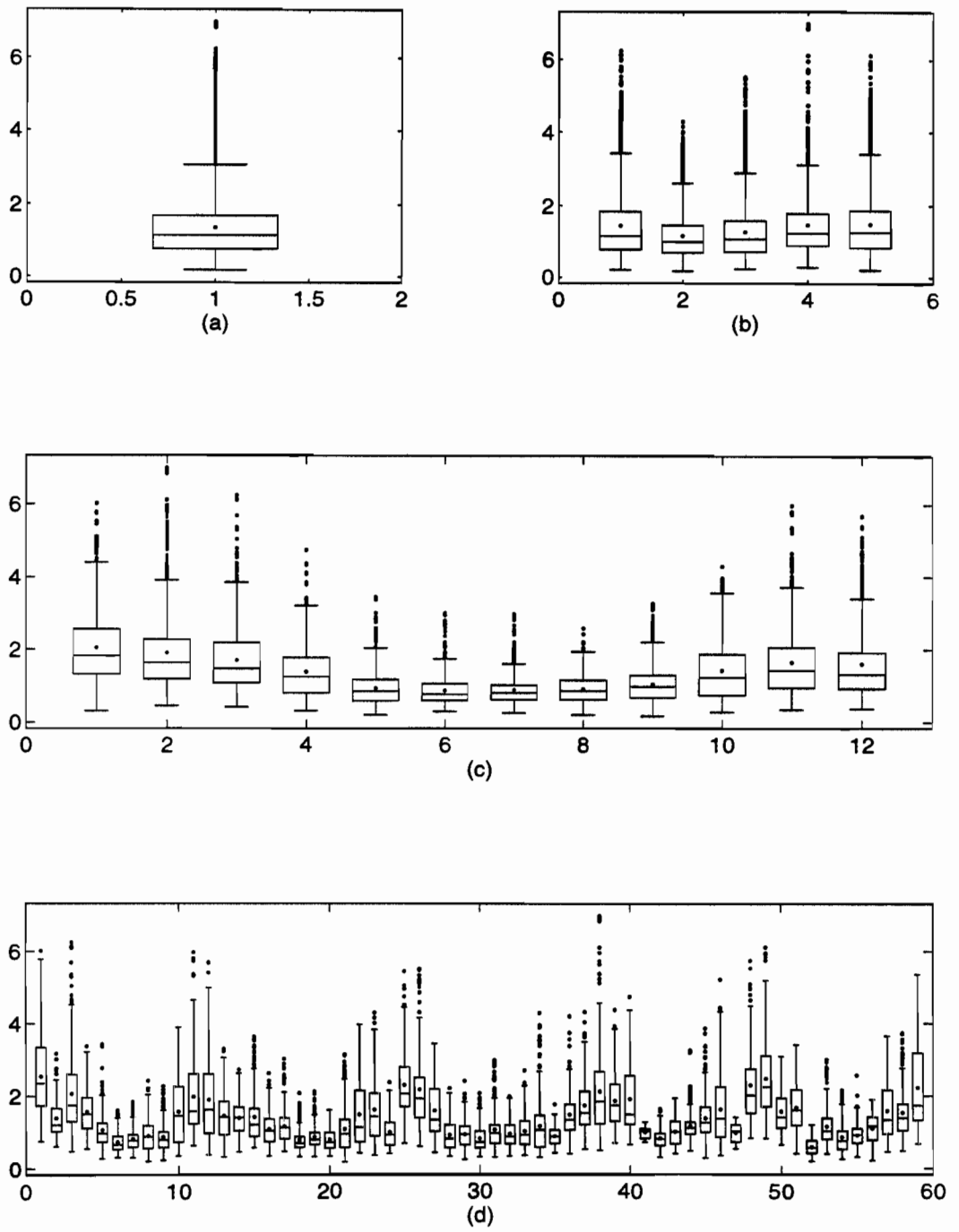


Figura 2: Box-plots de: a) toda la serie; b) datos de cada año; c) datos de cada mes; d) datos de todos los meses.

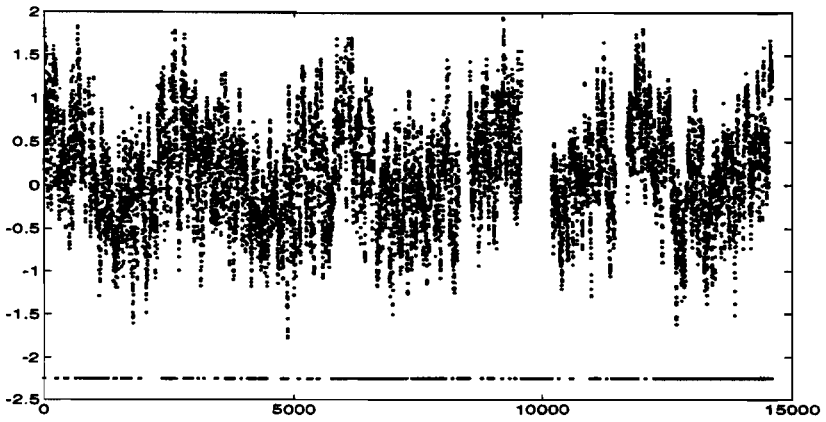


Figura 3: Serie en logaritmos de altura significativa de ola.

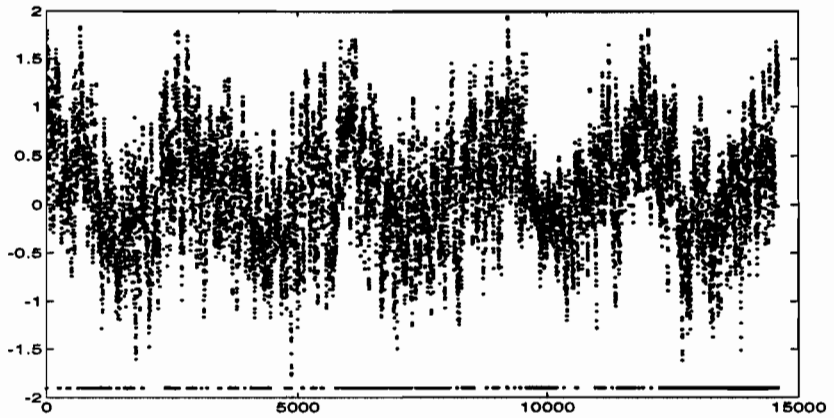
orden $s = 2920$, con lo que se pierden las observaciones correspondientes a un año y todas aquellas diferencias que incluyan un dato no registrado. El número de observaciones perdidas es tal que se hace imposible la estimación de las funciones de autocerrelación simple y parcial. Para evitar este problema proponemos: 1) interpolar la serie sustituyendo cada dato faltante por el logaritmo de la media de todas las observaciones registradas el mismo día y a la misma hora en cada año; 2) identificar y estimar un modelo ARIMA; 3) interpolar nuevamente la serie sustituyendo cada dato por la esperanza condicionada del dato faltante cuando el modelo es conocido. El resultado de la interpolación inicial se muestra en los gráficos de la figura **4**.

2.1. Identificación y estimación de un modelo ARIMA

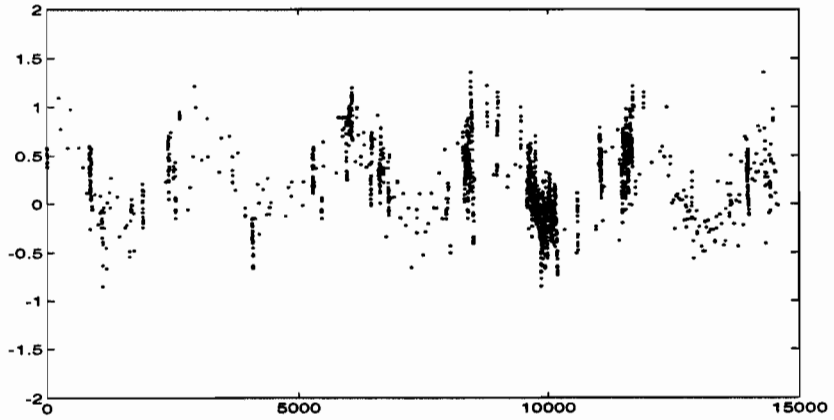
La expresión general de un modelo ARIMA para una serie temporal z_t es

$$(1-B)^d \phi(B) z_t = \theta(B) a_t, \tag{2.1}$$

donde $\phi(B) = (1 - \phi_1 B - \ldots - \phi_p B^p)$ y $\theta(B) = (1 - \theta_1 B - \ldots - \theta_q B^q)$ son los polinomios autorregresivo y media móvil respectivamente, d el número de raíces unitarias y B el operador de retardos ($Bz_t = z_{t-1}$). Las perturbaciones a_t son un proceso gaussiano de ruido blanco con varianza σ^2. Si suponemos que las raíces de θ están fuera del círculo



(a)



(b)

Figura 4: a) serie con interpolación anual; b) datos interpolados anualmente.

unidad la serie es invertible y podemos expresar (2.1) en forma autorregresiva

$$\pi(B)z_t = a_t,$$

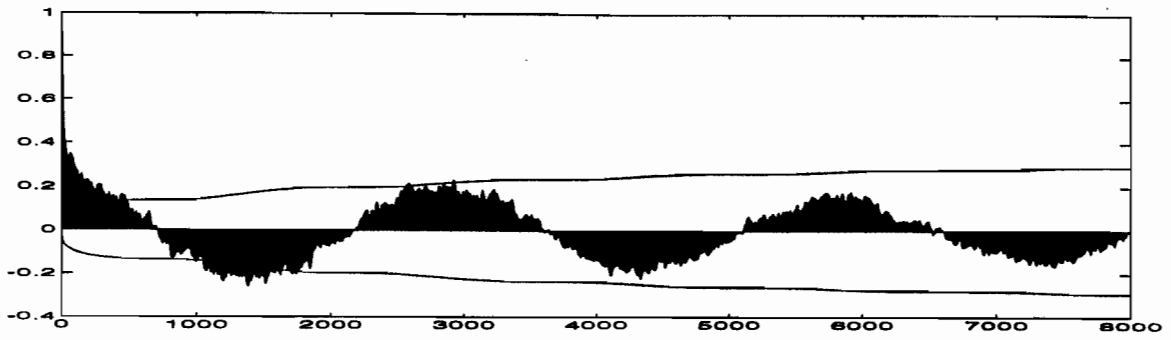
donde $\pi(B) = (1-B)^d\phi(B)/\theta(B) = (1 - \pi_1 B - \pi_2 B^2 \ldots)$.

La construcción de un modelo ARIMA (2.1) consiste en determinar el número de raíces unitarias del polinomio y los órdenes de los polinomios autorregresivo y media móvil, tanto para la parte regular como estacional. Una vez los órdenes son establecidos se estima el modelo y se calculan los residuos. El análisis de los residuos indica si el modelo seleccionado es el adecuado. Tanto la ACF como la PACF muestrales son las herramientas que habitualmente se usan para identificar los posibles modelos que se estiman, la selección de un modelo final la haremos basándonos en los siguientes criterios: parquedad en el número de parámetros, ausencia de estructura en la estimación de la ACF y la PACF de los residuos, varianza residual ($\hat{\sigma}^2$) y el R^2 del ajuste.

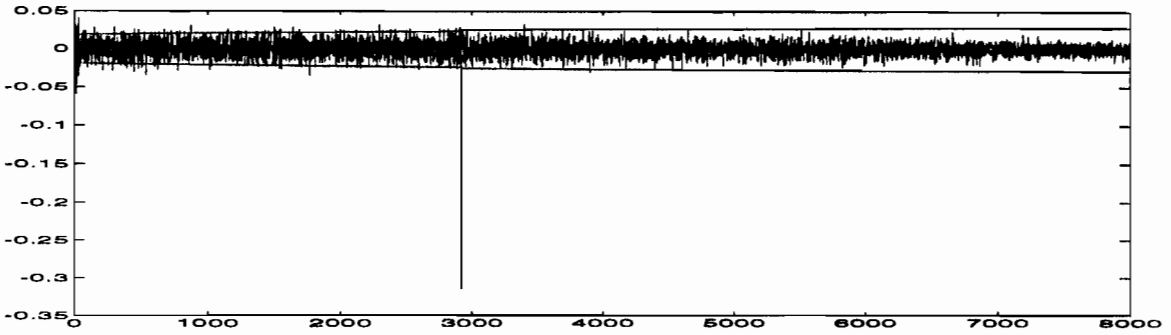
La ACF muestral de la serie interpolada (ver figura **5**a) muestra el comportamiento no estacionario de la serie tanto en la parte regular como estacional. En consecuencia, se toman diferencias regular y estacional y se estiman la ACF y la PACF para la serie diferenciada (ver figuras **5**b, **5**c y **5**d). Se identifican cuatro modelos ARMA$(p,q)(p_8,q_8)(p_s,q_s)$ estacionales para la serie diferenciada que junto con los estimadores de los parámetros, los estadísticos t, la varianza residual y el R^2, se recogen en la tabla **1**. Tanto la ACF como la PACF de los cuatro modelos no presentan estructura. El número de paramétros es bastante elevado en el modelo 1, por lo que este modelo se descarta. Como el resto de criterios no permiten discriminar claramente entre los modelos 2, 3 y 4 al ser los tres parámetros del modelo 2 significativos, seleccionamos el modelo MA(1)AR$_8$(1)MA$_s$(1).

2.2. Interpolación óptima

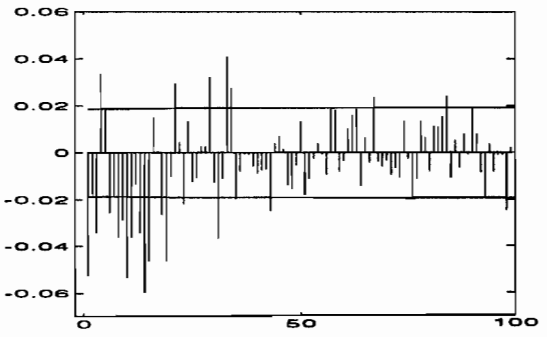
Para series infinitas y no estacionarias cuando el modelo ARIMA es conocido la esperanza condicionada de la observación faltante es un estimador óptimo, en el sentido de que minimiza el error cuadrático medio. Brubacher y Wilson (1976) demuestran que la expresión del estimador depende únicamente de la serie observada y de la función de autocorrelación del proceso dual, introducida por Cleveland (1972).



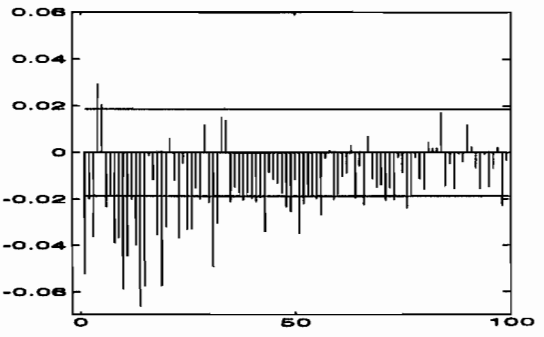
(a)



(b)



(c)



(d)

Figura 5: Interpolación anual: a) ACF de la serie; b) y c) ACF de la serie diferenciada; d) PACF de la serie diferenciada.

	Modelo	parám.	parám. estim.	estad. t	$\hat{\sigma}^2$	R^2
1	AR(16)$MA_s(1)$	ϕ_1	-0.0785	-8.48	0.0455	0.854
		ϕ_2	-0.0304	-3.28		
		ϕ_3	-0.0448	-4.84		
		ϕ_4	0.0179	1.94		
		ϕ_5	0.0063	0.68		
		ϕ_6	-0.0265	-2.87		
		ϕ_7	-0.0343	-3.72		
		ϕ_8	-0.0504	-5.48		
		ϕ_9	-0.0524	-5.68		
		ϕ_{10}	-0.0587	-6.36		
		ϕ_{11}	-0.0493	-5.33		
		ϕ_{12}	-0.0242	-2.62		
		ϕ_{13}	-0.0423	-4.57		
		ϕ_{14}	-0.0705	-7.62		
		ϕ_{15}	-0.0615	-6.63		
		ϕ_{16}	-0.0094	-1.01		
		θ_1^s	0.5363	55.84		
2	MA(1)$AR_8(1)MA_s(1)$	θ_1	0.0658	7.1	0.0465	0.851
		ϕ_1^8	-0.0434	-4.6		
		θ_1^s	0.5349	55.9		
3	MA(1)$MA_s(1)$	θ_1	0.0624	6.76	0.0465	0.851
		θ_1^s	0.5338	55.81		
4	$MA_s(1)$	θ_1^s	0.5345	55.69	0.0467	0.850

Table 1: Modelos ARIMA para la interpolación inicial.

Si suponemos que la varianza de las perturbaciones es $\sigma^2 = 1$, el filtro que se obtiene para interpolar un dato faltante en el instante $t = T$ es

$$\hat{z}_T = -\sum_{k=1}^{\infty} \rho_k^D z_{T-k} - \sum_{k=1}^{\infty} \rho_k^D z_{T+k}, \tag{2.2}$$

siendo $\rho_k^D = \rho_{-k}^D = \sigma_D^{-2} \sum_{j=0}^{\infty} \pi_j \pi_{j+k}$, la correlación de orden k del proceso dual. El coeficiente π_0 se define como $\pi_0 = -1$ y $\sigma_D^2 = \sum_{j=0}^{\infty} \pi_j^2$.

Si sólo observamos la serie hasta z_{T+n} y para algún índice $k > n$ el correspondiente coeficiente π_k del polinomio autorregresivo es positivo, la expresión del filtro óptimo varía en función de que se incorpore la corrección debida a la proximidad del dato faltante al final de la serie.

Para series finitas con observaciones faltantes próximas al final de la serie Maravall y Peña (1992) calculan el estimador óptimo del dato faltante. Este estimador corrige los pesos que definen la ponderación de cada observación en el filtro (2.2) y tiene la expresión

$$\hat{z}_{T,n} = -\sum_{k=1}^{\infty} \rho_{k,n}^D z_{T-k} - \sum_{k=1}^{n} \rho_{-k,n}^D z_{T+k}, \tag{2.3}$$

siendo $\rho_{k,n}^D = \sigma_{D,n}^{-2} \sum_{j=0}^{n} \pi_j \pi_{j+k}$ para $k \leq 1$ y $\rho_{-k,n}^D = \sigma_{D,n}^{-2} \sum_{j=0}^{n-k} \pi_j \pi_{j+k}$ para $k = 1, 2, \ldots, n$. En este caso $\sigma_{D,n}^2 = \sum_{j=0}^{n} \pi_j^2$.

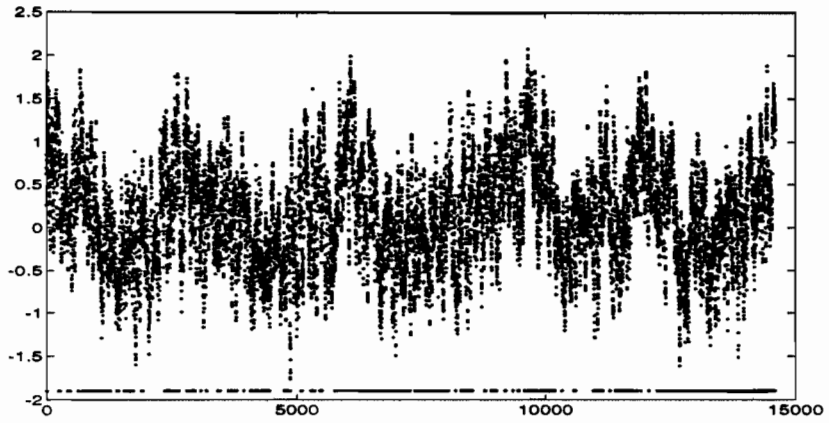
La expresión vectorial para el caso general en el que exista un vector de observaciones faltantes en distintos momentos del tiempo depende igualmente de la función de autocorrelación dual. El estimador que utilizamos para interpolar la serie de altura de ola z_t que incluye un vector de observaciones faltantes se puede encontrar en Maravall y Peña (1992) y el modelo a partir del cual se extraen los coeficientes del polinomio $\pi(B)$ es el discutido en el apartado anterior:

$$(1 + 0.04B^8)(1 - B)(1 - B^{2920})z_t = (1 - 0.06B)(1 - 0.53B^{2920})a_t.$$

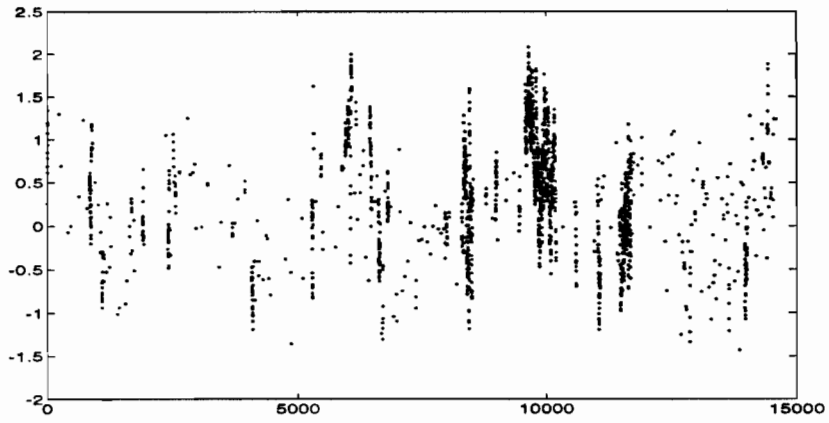
La serie interpolada óptimamente se muestra en el gráfico de la figura **6**.

3. MODELIZACIÓN DE LA SERIE DE OLEAJE

Reconstruida la serie de alturas de ola mediante interpolación óptima, el siguiente paso



(a)



(b)

Figura 6: a) serie con interpolación óptima; b) datos interpolados óptimamente.

es modelizar la serie. Para ello se proponen dos vías. La primera modelización consiste en ajustar un modelo ARIMA estacional a los datos siguiendo la misma metodología que se aplicó en la sección 2.

La segunda modelización difiere en el tratamiento de los efectos de la climatología en la altura de ola. Los ciclos de la serie son variables debido a que las estaciones meteorológicas no siempre llegan en las misma fechas del calendario. Es obvio que hay inviernos o veranos que se adelantan o atrasan. Este hecho no provocaría diferencias muy significativas en el componente estacional si la serie estuviese medida en periodos más largos, como meses o trimetres por ejemplo. Sin embargo, en la serie z_t se recogen los registros de altura de ola cada tres horas y el desfase se aprecia notablemente. Una simple diferencia estacional en este caso puede no ser la mejor forma de extraer el componente no estacionario estacional. Si suponemos que el modelo que sigue la serie es

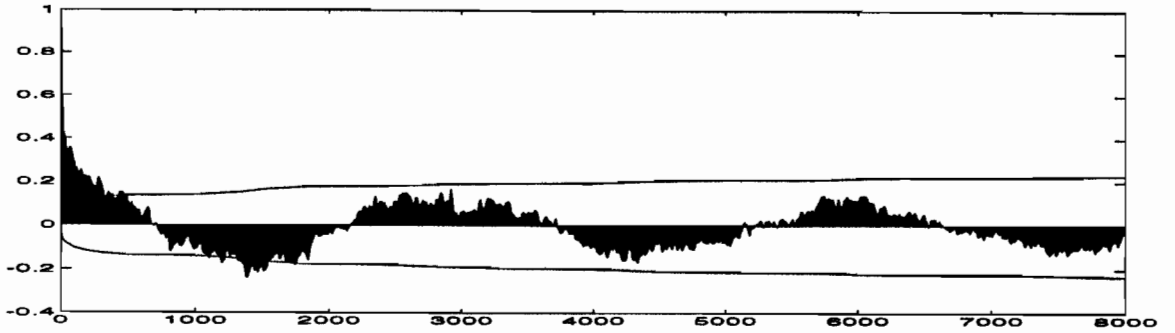
$$z_t = c(t) + x_t \qquad t = 1, \ldots, N,$$

donde $c(t)$ es una función del tiempo periódica, con periodo igual a un año, y x_t es un proceso estocástico que sigue un modelo ARIMA regular, la forma de eliminar los efectos estacionales que proponemos es realizar un suavizado de la serie mediante una regresión no paramétrica de los datos observados frente al tiempo.

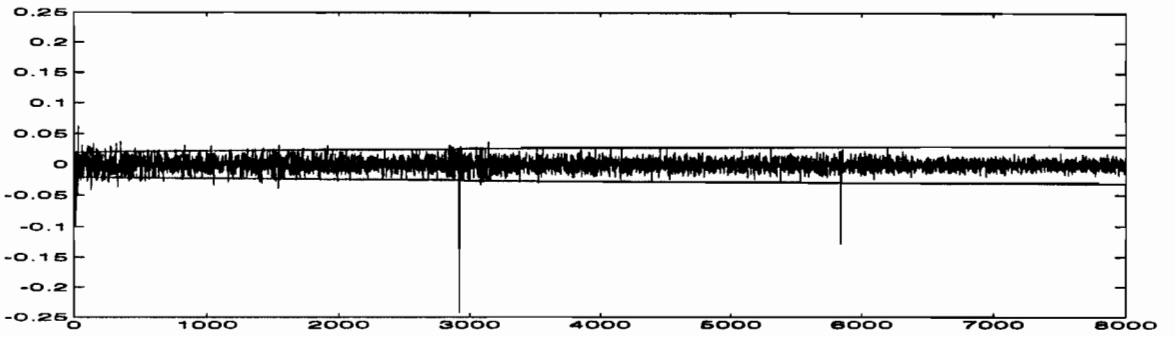
3.1. Modelos lineales

La modelización ARIMA de la serie interpolada óptimamamente z_t presenta cambios sustanciales con respecto a la que se realizó en la sección 2 para la serie reconstruida con la interpolación inicial. En la figura **7** se presentan estimaciones de la ACF y la PACF correspondientes a la serie interpolada óptimamente. Comparando estos gráficos con los de la figura **5** se aprecian diferencias significativas: 1) la correlación de orden 1 es positiva para la serie con interpolación óptima mientras que antes era negativa y 2) aparece una correlación negativa de orden 5840 ($2s$), que no estaba presente en el correlograma de la serie interpolada inicialmente.

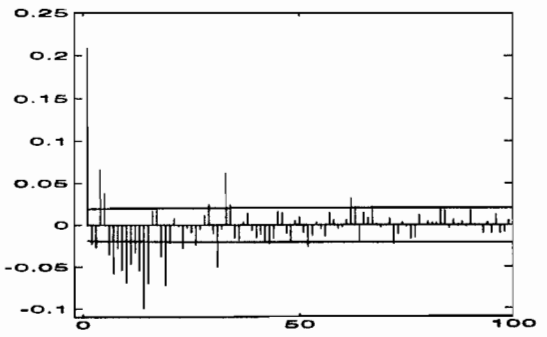
No es sorprendente la aparición de diferencias entre los correlogramas estimados a partir de las interpolaciones inicial y óptima. Si el modelo que se supone para intepolar



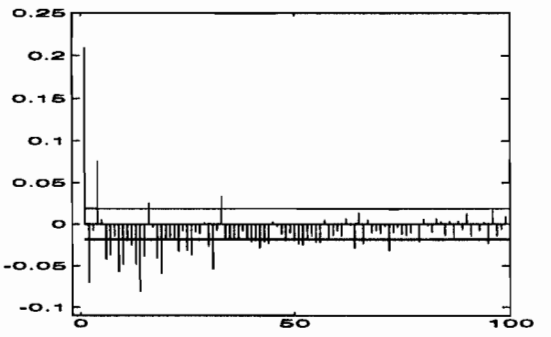
(a)



(b)



(c)



(d)

Figura 7: Interpolación óptima: a) ACF de la serie; b) y c) ACF de la serie diferenciada; d) PACF de la serie diferenciada.

óptimamente es cercano al verdadero modelo que sigue la serie, entonces esas diferencias ponen de manifiesto que la interpolación inicial tiene implicaciones perversas en la modelización. Sin embargo, existe una mayor evidencia a favor de que el correlograma cuando se interpola óptimamente es el que refleja la verdadera estructura de correlación de la serie.

En primer lugar, el correlograma obtenido a partir de la serie interpolada óptimamente es más acorde con la información que aportan los datos observados. La estimación del coeficiente de correlación de orden 1 con la interpolación óptima es $\hat{\rho}_1 = 0.2086$, mientras que con la estimación inicial es $\hat{\rho}_1 = -0.0519$. Comparemos estas estimaciones con un estimador del coeficiente de correlación ρ_1 independiente del método de interpolación. Este tercer estimador de ρ_1 se calcula a partir de secuencias de observaciones que en dos años consecutivos no incluyan datos faltantes. Usando todas las secuencias con más de 15 datos se estima $\hat{\rho}_1 = 0.1988$.

Por otro lado, la ACF y PACF que se estiman para una segunda interpolación de los datos (con el modelo que se identifica y estima a partir de la primera interpolación óptima) son prácticamente iguales a los proporcionados por la primera interpolación óptima.

Los modelos identificados para la serie z_t con interpolación óptima son los que se recogen en la tabla **2**. También se presentan la estimaciones de los parámetros, la varianza residual y el R^2. Teniendo en cuenta la parquedad en el número de parámetros, el modelo más satisfactorio es el $\mathrm{MA}(1)\mathrm{AR}_8(1)\mathrm{AR}_s(1)$

$$(1+0.03B^8)(1+0.53B^{2920}+0.46B^{5840}+0.19B^{8760})(1-B)(1-B^{2920})z_t = (1+0.17B)a_t.$$

3.2. Modelos mixtos

El suavizado de la serie se ha realizado mediante un estimador no paramétrico de la regresión tipo núcleo (véase Härdle, 1990)

$$\hat{c}(t) = \frac{\sum_{s=1}^{N} z_s K\left(\frac{t-s}{v}\right)}{\sum_{s=1}^{N} K\left(\frac{t-s}{v}\right)} \qquad t = 1, \ldots, N,$$

	Modelo	parám.	parám. estim.	estad. t	$\hat{\sigma}^2$	R^2
1	AR(16)$AR_s(3)$	ϕ_1	0.1692	9.13	0.0216	0.939
		ϕ_2	-0.0047	-0.25		
		ϕ_3	-0.0481	-2.58		
		ϕ_4	0.0653	3.50		
		ϕ_5	0.0042	0.23		
		ϕ_6	-0.0108	-0.58		
		ϕ_7	-0.0498	-2.67		
		ϕ_8	-0.0362	-1.94		
		ϕ_9	-0.0219	-1.17		
		ϕ_{10}	-0.0358	-1.92		
		ϕ_{11}	-0.0368	-1.97		
		ϕ_{12}	-0.0121	-0.65		
		ϕ_{13}	-0.0579	-3.11		
		ϕ_{14}	-0.0719	-3.86		
		ϕ_{15}	-0.0590	-3.15		
		ϕ_{16}	0.0042	0.22		
		ϕ_1^s	-0.5392	-28.29		
		ϕ_2^s	-0.4644	-25.72		
		ϕ_3^s	-0.2059	-12.09		
2	MA(1)$AR_8(1)AR_s(3)$	θ_1	-0.1751	-9.6	0.0224	0.937
		ϕ_1^8	-0.0384	-2.0		
		ϕ_1^s	-0.5359	-27.7		
		ϕ_2^s	-0.4635	-25.6		
		ϕ_3^s	-0.1998	-11.7		

Table 2: Modelos ARIMA para la interpolación óptima.

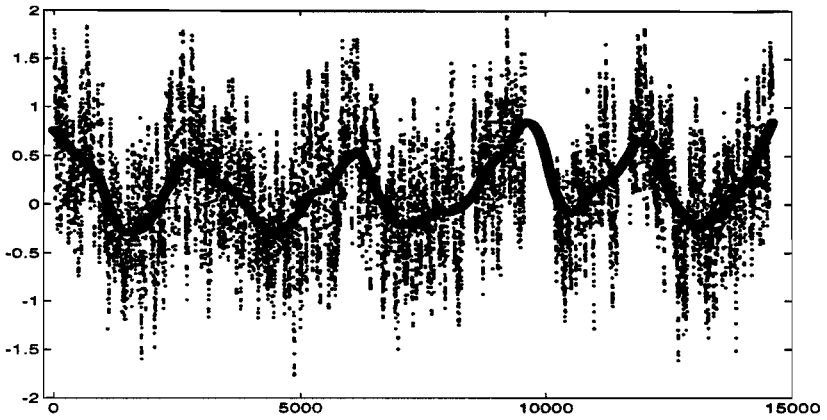
donde K es una función de densidad con varianza igual a 1, que se conoce como función *núcleo*. El parámetro v se denomina *parámetro de suavizado* o *ventana*. El estimador núcleo del valor de la función c en un punto t es una media ponderada de las observaciones registradas en instantes cercanos a s. El parámetro de suavizado v controla qué puntos se consideran cercanos a s, de manera que los instantes más próximos a s tendrán pesos tanto mayores cuanto menor sea v. Así, si v es grande el estimador de $c(t)$ es una función suave, mientras que si v es excesivamente pequeño entonces $\hat{c}(t)$ mantendrá parte de la variabilidad presente en la muestra. La elección del parámetro v es crucial en la estimación no paramétrica de la regresión. Sin embargo, el tipo de núcleo K afecta menos a la estimación. Se ha seleccionado el núcleo de *Epanechnikov*, que presenta algunas propiedades de optimalidad (véase Silverman, 1986, pag. 42). La elección de la ventana se hace de modo que se cumplan en la medida de lo posible los objetivos perseguidos con el suvizado de la serie: por un lado, que la estimación de x_t,

$$\hat{x}_t = z_t - \hat{c}(t),$$

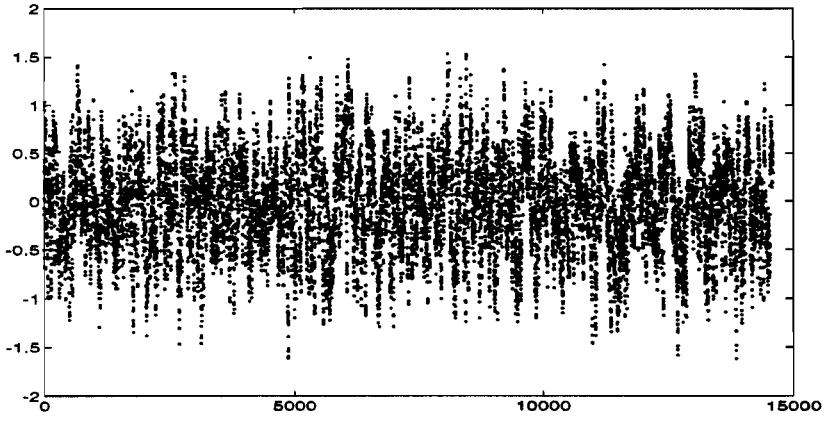
no contenga estructura estacional y, por otro, que $\hat{c}(t)$ sea próxima a una función periódica. El primer objetivo se logra tomando ventanas v suficientemente pequeñas, mientras que el segundo precisa de ventanas amplias. Así, el compromiso entre ambos criterios nos conduce a elegir como parámetro de suavizado el valor $v = 480$. Esto significa que en la estimación de $c(t)$ intervienen observaciones que distan del instante t hasta cuatro meses.

La estimación de $c(t)$ se ha realizado a partir de los datos z_t sin incluir las observaciones faltantes. En la figura **8**a se muestra el resultado de la estimación no paramétrica y en la figura **8**b se presentan los valores $\hat{x}_t$ para toda la serie interpolada óptimamente z_t.

A partir de la estimación de la componente cíclica de z_t se ajusta un modelo ARIMA regular a la estimación del proceso $\hat{x}_t = z_t - \hat{c}(t)$. En la figura **9**a se representa el correlograma de $\hat{x}_t$ para los 8000 primeros retardos. Puede observarse cómo las correlaciones estacionales (retardos múltiplos de s=2920) no son significativas o están muy próximas a las bandas de confianza. En cualquier caso se considera que los efectos estacionales han sido suficientemente mitigados mediante el suavizado y, por tanto, en $\hat{x}_t$ se modeliza únicamente la estructura regular. En ese mismo gráfico se aprecia evidencia de no



(a)



(b)

Figura 8: a) Estimación no paramétrica de la serie z_t; b) serie $\hat{x}_t = z_t - \hat{c}(t)$.

estacionariedad por lo que se opta por tomar una diferencia regular.

En las figuras **9**b y **9**c se muestran los 100 primeros retardos de la ACF y PACF muestral de la serie $\hat{x}_t$ diferenciada. Los modelos identificados a partir de estos gráficos son los incluidos en la tabla **3**, junto con las estimaciones de los parámetros, la varianza residual y el valor de R^2. Si seleccionamos el modelo con el criterio de parquedad en el número de parámetros, el modelo más apropiado es el $\mathrm{MA}(1)\mathrm{AR}_8(1)$ y, por tanto, para z_t es

$$(1 + 0.02B^8)(1 - B)(z_t - \hat{c}(t)) = (1 + 0.2B)a_t.$$

Teniendo en cuenta otros criterios para seleccionar entre modelos no hay una evidencia muy fuerte para seleccionar ninguno de los dos. La decisión final dependerá por tanto de qué modelo proponga mejores predicciones.

4. PREDICCIÓN A CORTO PLAZO

En esta sección se presenta un estudio comparativo de los modelos ARIMA estacionales y mixtos que se estiman en la sección 3. El objetivo es seleccionar el modelo que genera mejores predicciones de los datos futuros. El horizonte de predicción que interesa a las autoridades portuarias es de dos días, lo que equivale a 16 periodos adelante.

La predicción con modelos ARIMA se realiza de la forma habitual (Box y Jenkins, 1976). Cuando se utiliza un modelo mixto, la predicción es la suma del predictor de la función que representa el ciclo $c(t)$, más la predicción que se haga del proceso x_t mediante modelización ARIMA. La predicción del componente $c(t)$ se realiza usando como predictor del ciclo anual la media de los ciclos completos estimados mediante regresión no paramétrica

$$\hat{c}_N(N + t) = \frac{1}{A}\sum_{i=1}^{A} \hat{c}(t + (i - 1)s) \qquad t = 1, \ldots, s, \tag{4.4}$$

donde A es el número de años completos en la serie. Para evitar el efecto de los extremos de la serie en la estimación, en lugar del predictor (4.4) se sugiere el siguiente:

$$\hat{c}_N(N + t) = \frac{1}{A - 1}\sum_{i=1}^{A} \hat{c}(t + (i - 1)s) I_{[a_1, a_2]}(t + (i - 1)s) \qquad t = 1, \ldots, s, \tag{4.5}$$

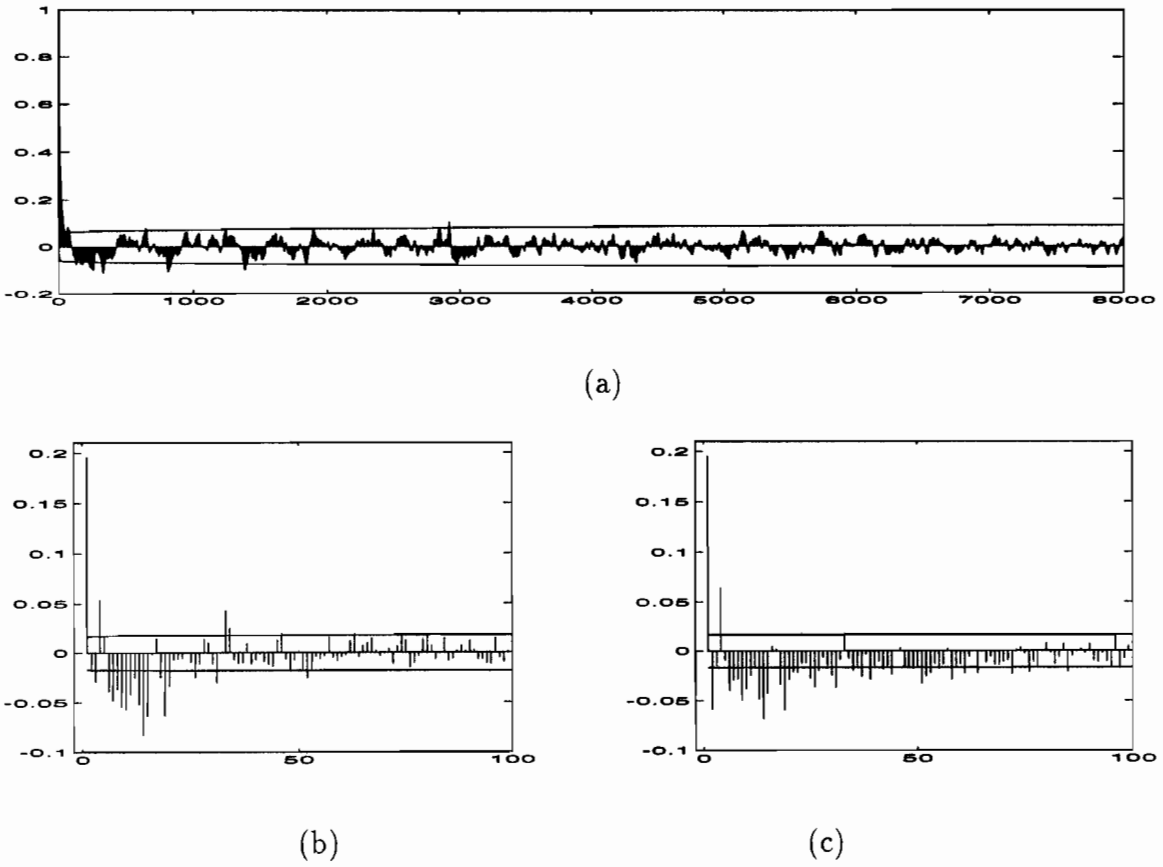


Figura 9: Serie $\hat{x}_t = z_t - \hat{c}(t)$: a) ACF de la serie; b) ACF de la serie diferenciada; c) PACF de la serie diferenciada.

	Modelo	parám.	parám. estim.	estad. t	$\hat{\sigma}^2$	R^2
1	AR(16)	ϕ_1	0.1940	23.43	0.0192	0.921
		ϕ_2	-0.0566	-6.72		
		ϕ_3	-0.0353	-4.19		
		ϕ_4	0.0570	6.77		
		ϕ_5	-0.0065	-0.77		
		ϕ_6	-0.0382	-4.53		
		ϕ_7	-0.0299	-3.55		
		ϕ_8	-0.0229	-2.72		
		ϕ_9	-0.0426	-5.06		
		ϕ_{10}	-0.0313	-3.71		
		ϕ_{11}	-0.0236	-2.80		
		ϕ_{12}	-0.0125	-1.48		
		ϕ_{13}	-0.0368	-4.37		
		ϕ_{14}	-0.0595	-7.06		
		ϕ_{15}	-0.0433	-5.14		
		ϕ_{16}	0.0048	0.58		
2	MA(1)AR_8(1)	θ_1	-0.2044	-25.2	0.0197	0.919
		ϕ_1^8	-0.0213	-2.5		

Table 3: Modelos ARIMA para la estimación de x_t.

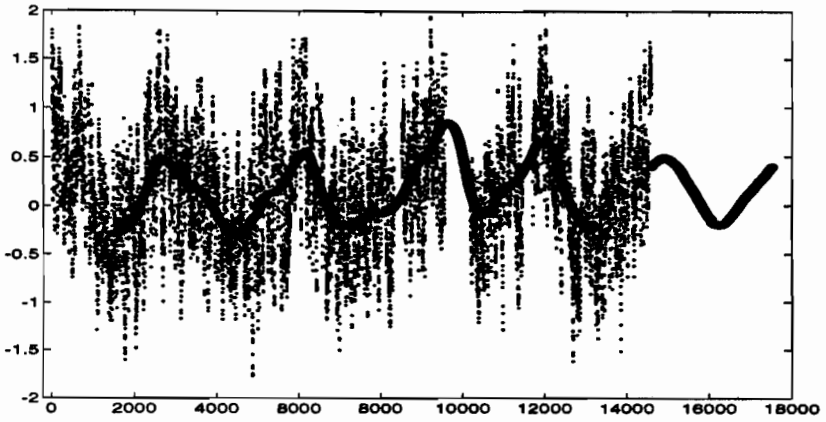
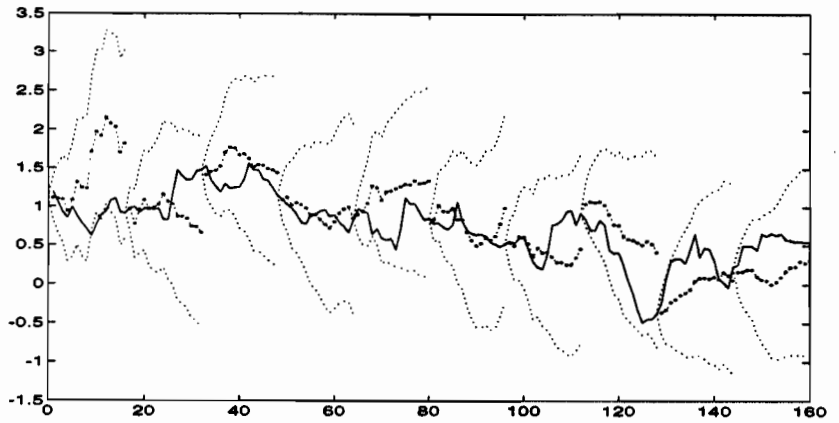


Figura 10: Estimación y predicción del ciclo anual

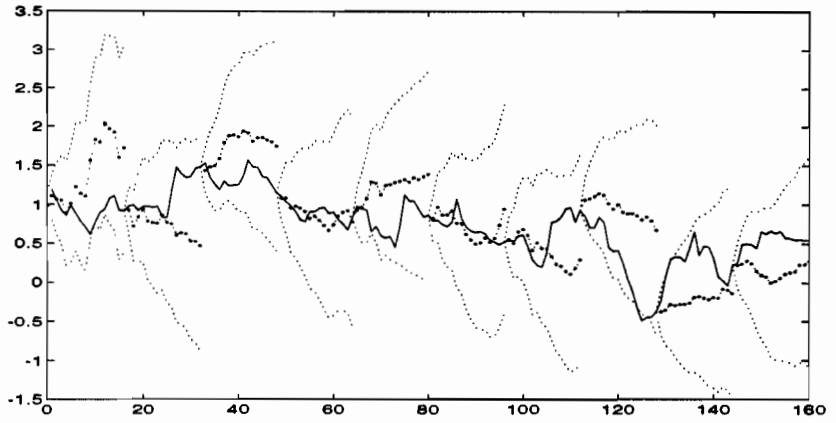
donde $a_1 = (1 + s/2)$ y $a_2 = ((A-1)s + s/2)$. El predictor (4.5) es el ciclo medio de la estimación de los cuatro ciclos completos que van desde el pricipio del segundo semestre de 1986 hasta el final de la primera mitad de 1990. En la figura **10** se muestra la estimación $\hat{c}(t)$ para los años observados y la predicción $\hat{c}_N(N+t)$ para el primer año fuera de la muestra.

Con los modelos considerados se realizan predicciones durante 20 días. Cada dos días se predicen los valores de la serie 16 períodos por delante. La primera predicción se realiza a las nueve de la noche del último día del año 1990, es decir, en el momento de la última observación de la serie modelizada. Como criterio de comparación entre modelos se toma el error cuadrático medio de predicción: la media de los cuadrados de las diferencias entre los valores reales y los predichos. Recordemos que se dispone de los datos reales registrados durante los veinte días en los que se realizan las predicciones. Los modelos utilizados para predecir han sido los cuatro modelos identificados y estimados en la sección 3.

Las figuras **11**a, **11**b, **12**a y **12**b muestran la serie real, las sendas de predicción y las bandas de confianza de las predicciones en los veinte días que se analizan. En la tabla **4** se recogen los errores cuadráticos medios de las predicciones para cada uno de los cuatro modelos cuando los horizontes de predicción van de uno a 16 periodos adelante. La última línea de esa tabla muestra los valores medios de los errores cuadráticos medios anteriores. Puede observarse que, salvo en la predicción a un paso, los modelos mixtos



(a)

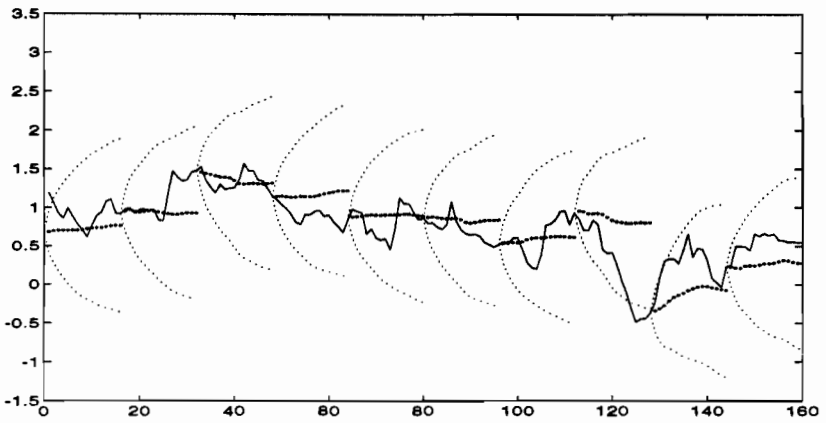


(b)

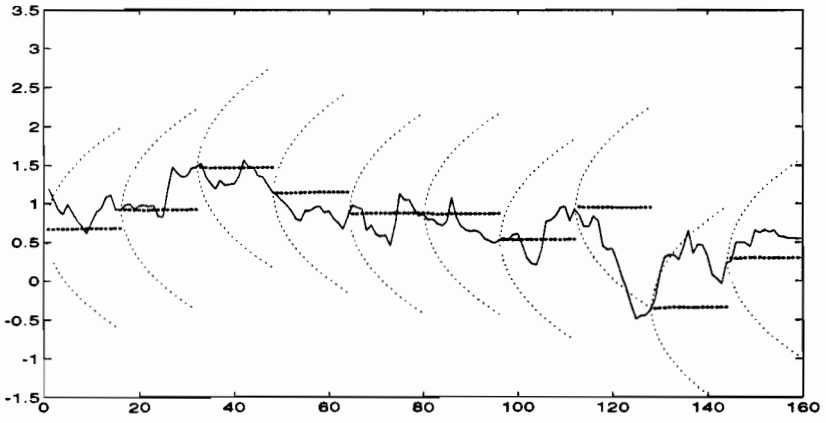
Figura 11: Predicción con modelos ARIMA estacionales: a) AR(16)$AR_s(3)$; b) MA(1)$AR_8(1)AR_s(3)$.

superan siempre a la modelización ARIMA pura. También se aprecia que la inclusión del polinomio autorregresivo de orden 16 en la parte regular de los modelos proporciona mejores predicciones. Los resultados obtenidos muestran que el modelo más apropiado para predecir con horizonte de predicción de dos días es el modelo mixto con componente lineal AR(16).

$$(1 - 0.19B + 0.05B^2 + \ldots + 0.004B^{16})(1 - B)(z_t - \hat{c}(t)) = a_t.$$



(a)



(b)

Figura 12: Predicción con modelos mixtos: a) modelo mixto con componente lineal **1** AR(16); b) modelo mixto con componente lineal **2** MA(1)AR_8(1).

	ARIMA estacional		**Mixto**	
	$MA(1)AR_8(1)AR_s(3)$	$AR(16)AR_s(3)$	$MA(1)AR_8(1)$	$AR(16)$
1 periodo	0.0086	0.0086	0.0325	0.0311
2 periodos	0.0513	0.0532	0.0492	0.0468
3 periodos	0.0666	0.0631	0.0678	0.0586
4 periodos	0.1152	0.1020	0.0754	0.0584
5 periodos	0.1233	0.0956	0.0786	0.0564
6 periodos	0.1882	0.1515	0.1153	0.0956
7 periodos	0.2215	0.1662	0.1333	0.0999
8 periodos	0.2518	0.1913	0.1736	0.1220
9 periodos	0.3034	0.2557	0.1384	0.0942
10 periodos	0.3497	0.2725	0.1842	0.1222
11 periodos	0.3730	0.2497	0.2420	0.1672
12 periodos	0.4240	0.3041	0.2654	0.1913
13 periodos	0.4194	0.2951	0.3067	0.2378
14 periodos	0.4230	0.3053	0.3001	0.2373
15 periodos	0.3976	0.2814	0.2914	0.2369
16 periodos	0.4185	0.2939	0.2980	0.2258
Total	0.2584	0.1931	0.1720	0.1301

Table 4: Error cuadratico medio de predicción con modelos ARIMA estacionales y modelos mixtos.

5. CONCLUSIONES

El modelo que finalmente se propone para predecir la altura significativa de ola a corto plazo es un modelo mixto en el que a la parte lineal, ajena al comportamiento cíclico de la serie, se le ajusta un modelo AR(16). Se ha comprobado que este es el modelo que incurre en menores errores cuadráticos medios de predicción en el periodo de veinte días utilizado para comparar los diferentes modelos propuestos.

Otra ventaja de este modelo es que el modelo ARIMA que se ajusta a la parte lineal no incluye diferencias de orden estacional y, en consecuencia, el cálculo de las predicciones requiere un tiempo de computación inferior al necesario para los modelos con diferencias estacionales.

La actualización del modelo propuesto debe hacerse periódicamente. Se propone que la reestimación del ciclo $c(t)$ se realice cada medio año y que la estimación de los parámetros del polinomio AR(16) se revise cada mes.

AGRADECIMIENTOS

Los autores agradecen a Daniel Peña y Juan Romo sus valiosos comentarios y a Obdulio Serrano y José Carlos Nieto el haberles facilitado la serie. Este trabajo ha sido parcialmente financiado por el Ente Público Puertos del Estado y por el proyecto PB93–0232 de la DGYCIT.

REFERENCIAS

Box, G.E.P. y Jenkins, G.M. (1970) *Time Series Analysis: Forecasting and Control.* Holden Day.

Brubacher, S.R. y Wilson, G.T. (1976) Interpolating time series with application to the estimation of holiday effects on electricity demand, *Applied Statistics,* 25, 2, 107–116.

Cleveland, W.P. (1972). The inverse autocorrelations of a time series and their applications, *Technometrics,* 14, 277–298.

GODA, Y. (1985). *Random Seas and Design of Maritime Structures,* Tokio: University of Tokio Press.

HÄRDLE, W. (1990). *Applied Non-parametric Regression,* Oxford University Press.

MARAVALL, A. Y PEÑA, D. (1992). Missing observations and additive outliers in time series models. *Advances in Statistical Analysis and Statistical Computing.* JAI Press (en prensa).

SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis,* Londres: Chapman and Hall.

SORENSEN, R.M. (1993). *Basic Wave Mechanics for Coastal and Ocean Engineers,* Nueva York: John Wiley.