# FORECASTING WITH MISSING DATA: APPLICATION TO A REAL CASE

Pedro Delicado* and Ana Justel**

*Department of Economics and Business, Universitat Pompeu Fabra

** Department of Mathematics, Universidad Autónoma de Madrid

**Abstract**

This paper presents a comparative analysis of linear and mixed models for short term forecasting of a real data series with a high percentage of missing data. Data are the series of significant wave heights registered at regular periods of three hours by a buoy placed in the Bay of Biscay. The series is interpolated with a linear predictor which minimizes the forecast mean square error. The linear models are seasonal ARIMA models and the mixed models have a linear component and a non linear seasonal component. The non linear component is estimated by a non parametric regression of data versus time. Short term forecasts, no more than two days ahead, are of interest because they can be used by the port authorities to notice the fleet. Several models are fitted and compared by their forecasting behavior.

*Key words:* Significant wave height; mean square error; linear interpolation; ARIMA models; nonparametric smoothing.

*AMS Classification:* 62M10, 62M20.

# 1. INTRODUCTION

The study of natural phenomena provides large data bases from measurements of physical magnitudes. In many cases, these measurements are registered in regular time intervals and it is possible to build time series, which present particular characteristics depending on the studied physical phenomenon and on the used instrumental.

In this paper we analyze the series of significant wave heights registered every three hours by a buoy located in the Bay of Biscay. Particular characteristics of the series are the large size (14608 data), the unusual periodicity ($s = 2920$ data, one year), the possibility of daily seasonality, and the large number of missing data due to data transmission failures and damages in the measuring devices. We discuss a methodology to analyze the series which can be applied to series with similar characteristics. The interest is in proposing a model for these series in order to carry out fast short term predictions, one or two days ahead. These forecast horizons are important for the port authorities since they can alert to the fleet of sea variations. For this reason we propose univariate models for the significant wave height series in spite of the fact that the predictions could improve with a structural model including information of temperatures, winds, etc. The analysis of the significant wave height series is divided in three stages: (1) missing data interpolation; (2) identification and estimation of appropriate models; and (3) model selection.

The amount of missing data (approximately a 13%) and the times when they appear make difficult the model identification and estimation. Most of the missing data are isolated and the intervals between them are relatively short, however there are also some long periods without registrations. Moreover, when a model includes information from the data in the previous years, the lack of only one data will prevent the model from predicting the values in the same instants of posterior years. Thus, our first task is to overcome the problems derived from the large number of missing data.

A natural missing data interpolation using the mean of all the data registered the same day in other years allows us to identify a model. With this initial model the optimal missing data interpolation procedure proposed by Maravall and Peña (1997) is used to complete the series. Then, we propose two different types of models for the optimally

interpolated series: on the one hand, seasonal ARIMA linear models following the Box-Jenkins methodology (Box and Jenkins, 1976), and on the other, mixed models with a non linear part for the seasonal component and a linear part for the regular dependence structure. We select the model which provides better out-of-sample short term forecasts by comparing the predictions with the data observed in the last month (not used in the estimation stage).

This article is organized as follows. Next subsection deals with the significant wave height definition and with the data description. Section 2 analyzes the problem of missing data interpolation. Section 3 discusses the two types of models: linear and mixed models. Section 4 compares the short term forecasts from the models. Finally, section 5 presents some final comments.

### 1.1. Significant wave height

The wave height is defined as the distance between the minimum wave value, *valley*, and the maximum value, *crest*. This distance is usually computed by using special instruments positioned in buoys. The instruments register the wave accelerations in short time intervals during a period of approximately 30 minutes. The aboard sensor integrates twice the series of accelerations to obtain the series $\eta_i$ of wave heights. The short term sea surface elevation is a stochastic stationary process in mean since it is assumed that the sea level is constant, and stationary in variance.

For many years the surge information came from visual registrations on ships in route and the human perception tends to overvalue the wave height. Therefore, in order to make compatible the historical information with the current automatic data, it is necessary to define a parameter which permits to join up both sources of information. The commonly used parameter is the mean value of the third higher heights, known as *significant wave height*.

The significant wave height $h$ can be approached by $4\sigma_\eta$, where the variance $\sigma_\eta^2$ is the integral of the spectral density $f$ of the series $\eta_i$, $\sigma_\eta^2 = \int f(\nu)d\nu$. The spectral density $f$ is computed by means of the fast Fourier transformation (FFT) and $h$ is obtained by

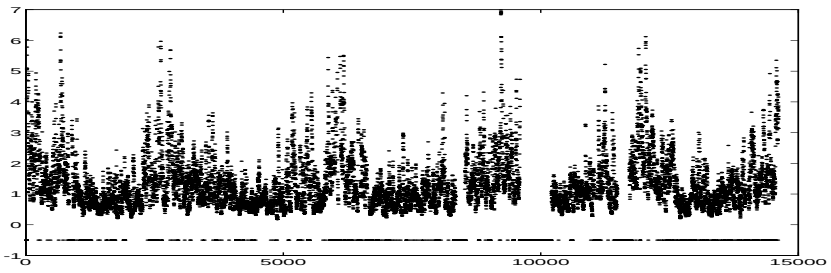



**Figure 1**: Significant wave height series and the missing data in the lower line.

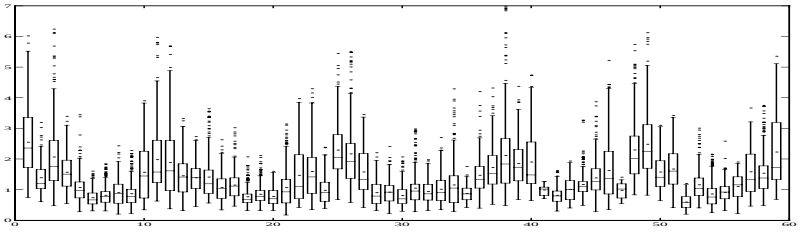



**Figure 2**: Series of box-plots with the every month data.

numerical integration of $f$. This procedure is repeated periodically and a series of values $h_t$ is recorded. For more details on the construction of the surge series see the books of Goda (1985) and Sorensen (1993).

The series $h_t$ we analyze here comes from the registrations in a buoy located near the Gijón coast (north of Spain). The acceleration is measured every 0.5 seconds in a total of 5120 instants, this means an observation period of about 42 minutes. These registrations are carried out every three hours producing the series of significant wave height. The available data are the 14608 significant wave heights recorded every three hours from 1-1-1986 to 1-31-1991, except 1871 missing data. Figure 1 shows the series and, in the lower line, the missing data. In Figure 2 we present a time series of box-plots constructed with the data in every month. We observe that data are asymmetric and more variable in the winter months. Due to the heteroskedasticity and asymmetry problems we transform the data. From now on, we work with the series $z_t$ of significant wave height logarithms, $z_t = \log h_t$.

## 2. MISSING DATA INTERPOLATION

The autocorrelation function (ACF) and the partial autocorrelation function (PACF) are the main tools used in the ARIMA model identification stage. However, the autocorrelation function estimation is difficult when there are missing data in the series. The number of available data to estimate each correlation coefficient $\rho_k$ decreases rapidly when $k$ raises.

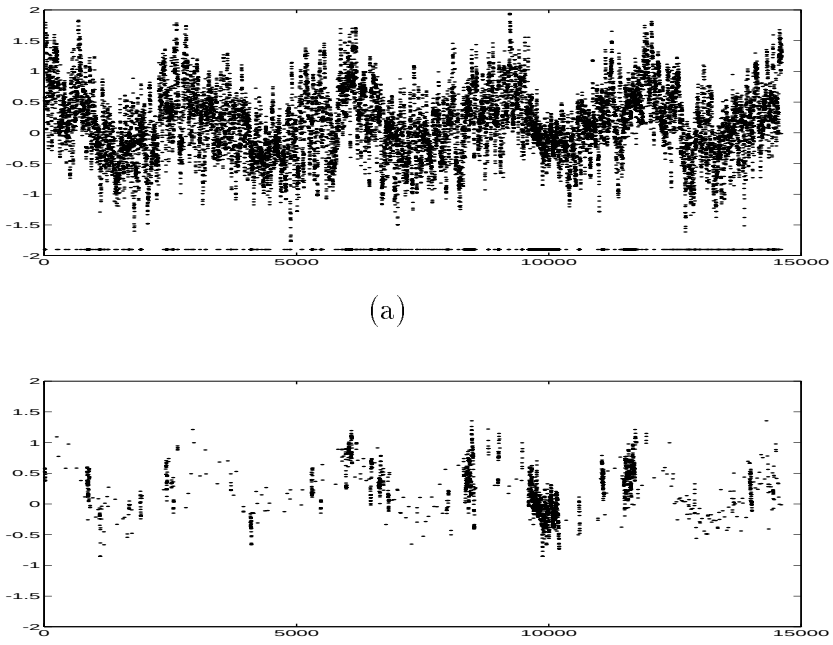
In spite of the absence of 1871 data in the series $z_t$ we could estimate its ACF since there are still many observations in this series. However, due to the non stationary behavior of the series it is necessary to transform $z_t$ with a regular and a seasonal differences (the last one is of order $s = 2920$). Then one year and all the differences including a non registered data are missed. The number of missing observations is such that it is impossible to estimate the simple and partial autocorrelation function for the differentiated series. To avoid this problem we propose the following procedure: 1) interpolate the series replacing each missing data by the logarithm of the mean of all the data registered the same day and at the same hour in every year (the result of this initial interpolation is shown in Figure 3); 2) identify and estimate an ARIMA model for this naive interpolation; and 3) interpolate again the original series substituting each missing data by its conditional expected value given the observations and assuming that the previous model is the actual data generator process (this will be called the *optimal interpolation*).

### 2.1. ARIMA models: identification and estimation.

A general ARIMA model expression for a time series $z_t$ is

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

where $\phi(B) = (1 - \phi_1 B - \ldots - \phi_p B^p)$ and $\theta(B) = (1 - \theta_1 B - \ldots - \theta_q B^q)$ are the autoregressive and moving average polynomials respectively, $\Phi(B^s) = (1 - \phi_1^s B^s - \ldots - \phi_P^s B^{Ps})$ and $\Theta(B^s) = (1 - \theta_1^s B^s - \ldots - \theta_Q^s B^{Qs})$ are the seasonal autoregressive and moving average polynomials respectively, $d$ and $D$ are the number of regular and seasonal required




(a)




(b)

**Figure 3:** a) Series with initial interpolation; b) interpolated data.

differences to achieve stationarity. The disturbances $a_t$ are assumed to follow a white noise gaussian process with variance $\sigma^2$.

The ARIMA model identification consists on determining the regular and seasonal orders of differentiation, and the degrees of the autoregressive and moving average lag polynomials for the regular and the seasonal parts. The final selection is usually based on several criteria: (a) parsimony in the number of parameters; (b) non structure in the residual sample ACF and PACF; (c) lower residual variance $\hat{\sigma}^2$; and (d) higher adjusted determination coefficient $R^2$ (for models with identical number of differences).

The sample ACF of the interpolated series shows the non stationary behavior of the series in the regular and seasonal parts (see Figure 4a). Therefore, we differentiate the series with a regular and a seasonal differences. Sample ACF and PACF for the differentiated series are shown in Figures 4b, 4c and 4d. We identify four seasonal ARMA$(p,q)(p_8,q_8)(p_s,q_s)$ models. They are displayed in Table 1, as well as the parameter

| Model | Parameters | Estimated parameters (t-value) |
|---|---|---|
| **(1)** AR(16)MA$_s$(1) | $\phi_1$ $\phi_2$ $\phi_3$ $\phi_4$ | -0.078 (-8.5) -0.030 (-3.2) -0.045 (-4.8) 0.017 ( 1.9) |
| $\hat{\sigma}^2 = 0.045$, $R^2 = 0.854$ | $\phi_5$ $\phi_6$ $\phi_7$ $\phi_8$ | 0.005 ( 0.7) -0.028 (-2.9) -0.034 (-3.7) -0.050 (-5.5) |
| | $\phi_9$ $\phi_{10}$ $\phi_{11}$ $\phi_{12}$ | -0.052 (-5.9) -0.087 (-6.4) -0.093 (-5.3) -0.029 (-2.6) |
| | $\phi_{13}$ $\phi_{14}$ $\phi_{15}$ $\phi_{16}$ | -0.042 (-4.6) -0.070 (-7.6) -0.061 (-6.6) -0.009 (-1.0) |
| | $\theta_1^s$ | 0.536 (55.8) |
| **(2)** MA(1)AR$_8$(1)MA$_s$(1) | $\theta_1$ | 0.066 (7.1) |
| $\hat{\sigma}^2 = 0.046$, $R^2 = 0.851$ | $\phi_1^8$ | -0.043 (-4.6) |
| | $\theta_1^s$ | 0.535 (55.9) |
| **(3)** MA(1)MA$_s$(1) | $\theta_1$ | 0.062 (6.8) |
| $\hat{\sigma}^2 = 0.046$, $R^2 = 0.851$ | $\theta_1^s$ | 0.534 (55.8) |
| **(4)** MA$_s$(1) | $\theta_1^s$ | 0.534 (55.7) |
| $\hat{\sigma}^2 = 0.046$, $R^2 = 0.850$ | | |

**Table 1**: ARIMA models for the initial interpolation with regular and seasonal differences, $s = 2929$.

estimates, the $t$ statistics, the residual variance and the $R^2$ values. Neither the ACF nor the PACF for the four models present structure. Model 1 is discarded because of the high number of parameters. The other criteria are not very useful to discriminate among the models 2, 3 and 4. We select the MA(1)AR$_8$(1)MA$_s$(1) because its three parameters are significant.

## 2.2. Optimal interpolation

When the ARIMA model is known, for infinite and non stationary series the conditional expectation of the missing data is the optimal predictor in the sense that it minimizes the mean square error of prediction. Brubacher and Wilson (1976) proved that this predictor only depends on the observed series and on the autocorrelation function of the dual process introduced by Cleveland (1972). When the series is only observed up to $z_{T+n}$ and any coefficient $\pi_k$ of $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 \ldots$ (the infinite autoregressive representation of the model (2.1), $\pi(B) z_t = a_t$) is positive for $k > n$, the optimal filter

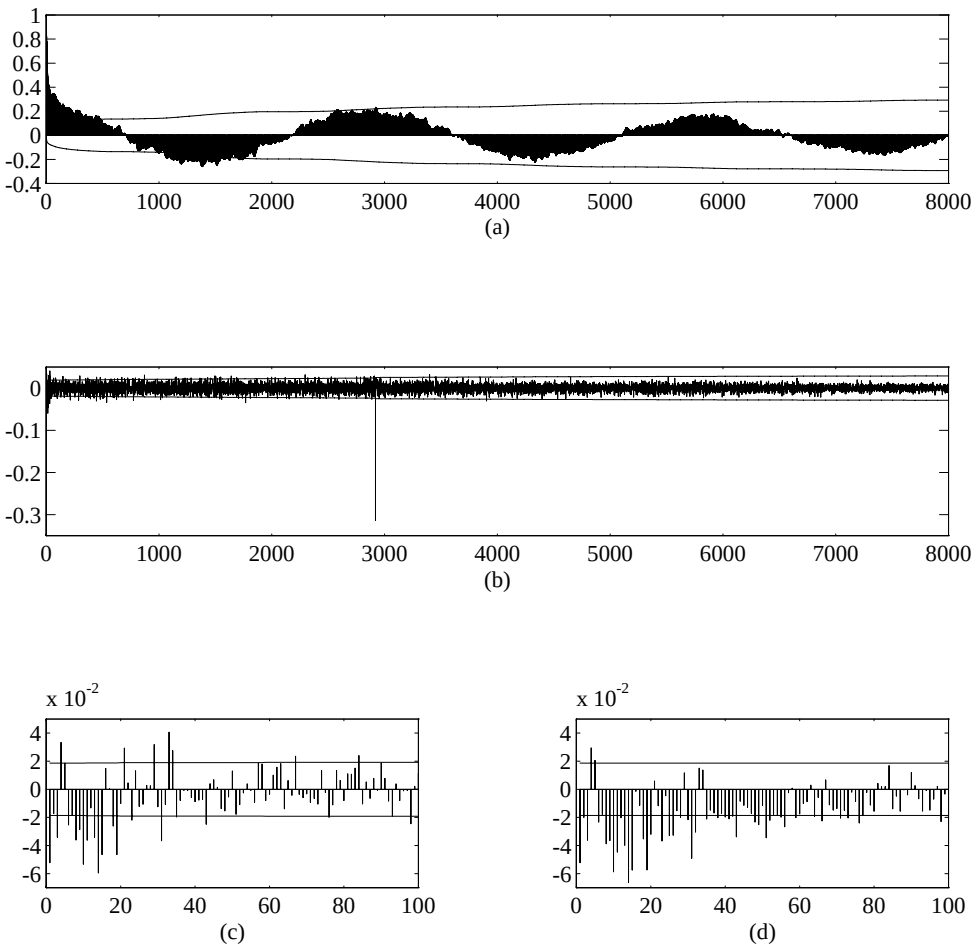



**Figure 4:** Initial interpolation: a) sample ACF of the series; b) and c) sample ACF of the differentiated series; d) sample PACF of the differentiated series.

must incorporate a correction for missing data near to the end of the series. Maravall and Peña (1997) obtained the optimal missing data predictor by correcting the weights of each observation. Assuming that $\sigma^2 = 1$, the predictor is given by

$$\hat{z}_{T,n} = -\sum_{k=1}^{\infty} \rho^D_{k,n} z_{T-k} - \sum_{k=1}^{n} \rho^D_{-k,n} z_{T+k}, \tag{2.2}$$

where , $\rho^D_{k,n} = \sigma^{-2}_{D,n} \sum_{j=0}^{n} \pi_j \pi_{j+k}$ for $k \leq 1$ and $\rho^D_{-k,n} = \sigma^{-2}_{D,n} \sum_{j=0}^{n-k} \pi_j \pi_{j+k}$ for $k = 1, 2, \ldots, n$ are the autocorrelations of the truncated dual process, the coefficient $\pi_0$ is defined as $\pi_0 = -1$ and $\sigma^2_{D,n} = \sum_{j=0}^{n} \pi_j^2$. .

The significant wave height series $z_t$ includes more than a single missing data. The optimal predictor for this general case also depends on the dual autocorrelation function in a similar way and its expression is given by Maravall and Peña (1997). The coefficients of $\pi(B)$ are computed from the estimated model:

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

The series with optimal interpolation is shown in Figure 5. We can appreciate the differences between the naive and the optimal approaches to interpolate the series comparing Figures 3 and 5. It seems that the variability of the original series is better captured by using the optimal interpolator.

## 3. MODEL IDENTIFICATION

We propose two alternative types of models for the optimally interpolated series. The first models are seasonal ARIMA models identified with the same methodology applied in the section 2. The second models differ from the former in the consideration of the climatological effects on the significant wave height. We observe that the low frequency cycles are variable because the meteorological stations not always arrive in the same dates of the calendar. It is obvious that there are winters or summers that anticipate or retard their presence. This fact would not cause very significant differences in the seasonal component if the series is measure in longer periods, like months or quarters for example.

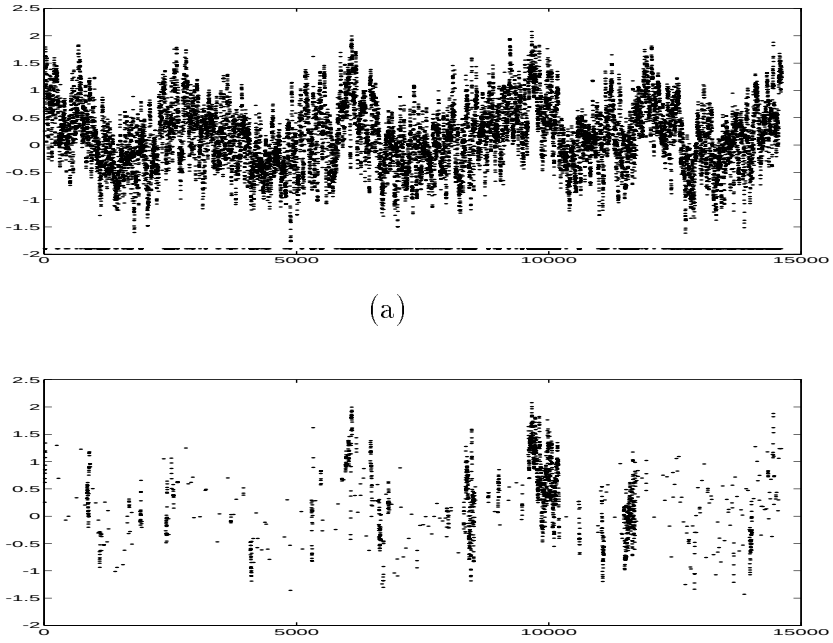



**Figure 5:** a) Series with optimal interpolation; b) interpolated data.

However, data in $z_t$ are collected every three hours and the disagreement between different years is notably appreciable. In this case a seasonal difference could not be the best way to eliminate the non stationary seasonal component. We propose to eliminate the seasonal effects by smoothing the series with a nonparametric regression of data versus time. Then we assume the model

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

where $c(t)$ is almost a periodic function with one year period and $x_t$ is a process following a regular ARIMA model.

### 3.1. Linear model

The ARIMA model selection for the series with optimal interpolation presents essential changes with respect to the work presented in section 2 for the series with the initial interpolation. In Figure 6 the sample ACF and PACF are displayed for the series with

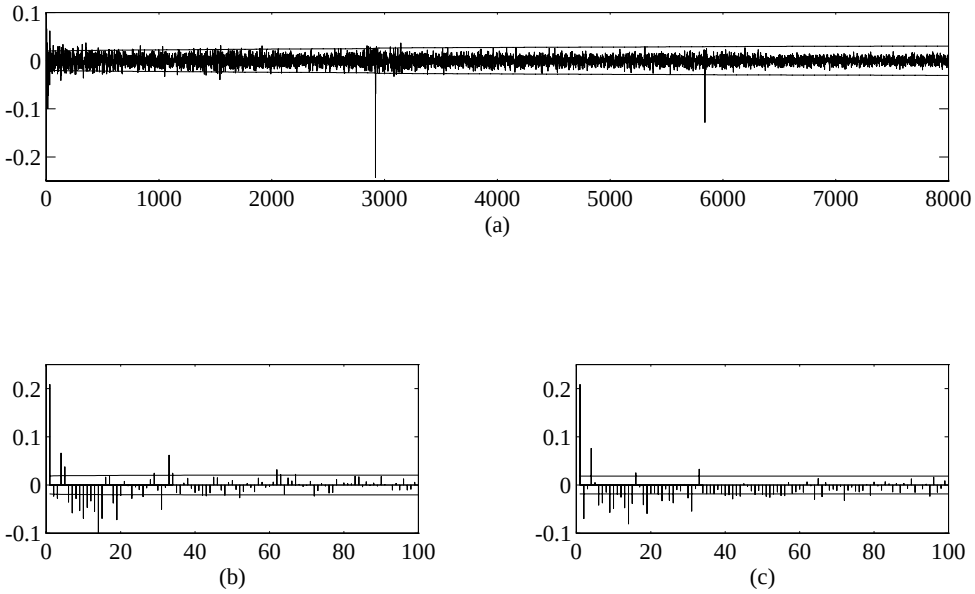



**Figure 6**: Optimal interpolation: a) and b) ACF of the differentiated series; c) PACF of the differentiated series.

optimal interpolation and a regular and a seasonal differences. We appreciate the next differences with respect to the ACF and PACF in Figure 4: 1) the first order autocorrelation coefficient is positive for the series with optimal interpolation while before was negative; and 2) there is a negative autocorrelation of order $2s = 5840$ which was not present before.

Some evidences confirm that the model used for the optimal interpolation is closer to the actual model than the model obtained from naive interpolation. Firstly, the ACF of the optimally interpolated series is more in accordance with the observed data information than the ACF of naively interpolated data. To support this comment, we compare the first order autocorrelation coefficient $\rho_1$ with an estimate that is independent of the interpolation method: it is computed with all the data sequences non including missing data in two consecutive years. Using all of these sequences with more than 15 data we obtain an estimation of 0.1988 for $\rho_1$, while with the optimal interpolation $\hat{\rho}_1 = 0.2086$ and with the initial estimation $\hat{\rho}_1 = -0.0519$. Additionally, we can iterate the process

| Model | Parameters | Estimated parameters (t-value) |
|---|---|---|
| **(1)** $\mathrm{AR}(16)\mathrm{AR}_s(3)$ | $\phi_1$ $\phi_2$ $\phi_3$ $\phi_4$ | 0.169 (9.1) -0.005 (-0.2) -0.048 (-2.6) 0.065 (3.5) |
| $\hat{\sigma}^2 = 0.022$, $R^2 = 0.939$ | $\phi_5$ $\phi_6$ $\phi_7$ $\phi_8$ | 0.004 ( 0.2) -0.011 (-0.6) -0.050 (-2.7) -0.036 (-1.9) |
| | $\phi_9$ $\phi_{10}$ $\phi_{11}$ $\phi_{12}$ | -0.022 (-1.2) -0.036 (-1.9) -0.037 (-2.0) -0.012 (-0.6) |
| | $\phi_{13}$ $\phi_{14}$ $\phi_{15}$ $\phi_{16}$ | -0.058 (-3.1) -0.072 (-3.9) -0.059 (-3.1) 0.004 (0.2) |
| | $\phi_1^s$ $\phi_2^s$ $\phi_3^s$ | -0.539 (-28.3) -0.464 (-25.7) -0.206 (-12.1) |
| **(2)** $\mathrm{MA}(1)\mathrm{AR}_8(1)\mathrm{AR}_s(3)$ | $\theta_1$ | -0.175 (-9.6) |
| $\hat{\sigma}^2 = 0.022$, $R^2 = 0.937$ | $\phi_1^8$ | -0.038 (-2.0) |
| | $\phi_1^s$ $\phi_2^s$ $\phi_3^s$ | -0.536 (-27.7) -0.463 (-25.6) -0.200 (-11.7) |

**Table 2:** ARIMA models for the series with optimal interpolation and a regular and a seasonal differences, $s = 2929$.

of optimal interpolation and interpolate again the series using as initial model the one identified from the first optimal interpolation. If we do that, the sample ACF and PACF of the series with second optimal interpolation are practically equal to those obtained in the previous step, displayed in Figure 4. Thus, the identified model after the first optimal interpolation is very near to the "fix point" model of that iterative process. The differences in ACF's and PACF's that we appreciate show that the initial interpolation might produce perverse implications in the model identification.

We present in Table 2 the identified models for the series $z_t$ with optimal interpolation. Considering the parsimony in the number of parameters, we select the $\mathrm{MA}(1)\mathrm{AR}_8(1)\mathrm{AR}_s(1)$ model:

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

### 3.2. Mixed models

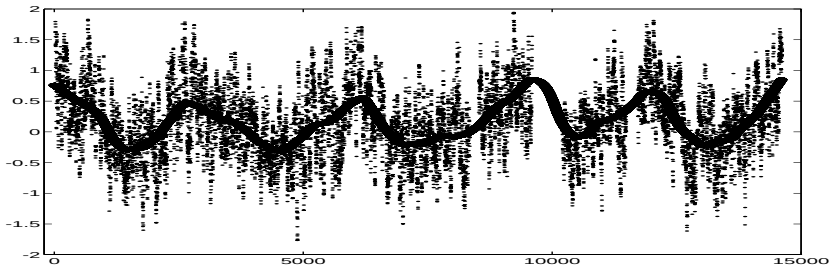
We propose to smooth the series with a nonparametric Nadaraya-Watson kernel estimate of the regression function (see, for instance, Härdle, 1990),

$$\hat{c}(t) = \frac{\sum_{i=1}^{N} z_i K\left(\frac{t-i}{b}\right)}{\sum_{i=1}^{N} K\left(\frac{t-i}{b}\right)}, \qquad t = 1, \ldots, N,$$

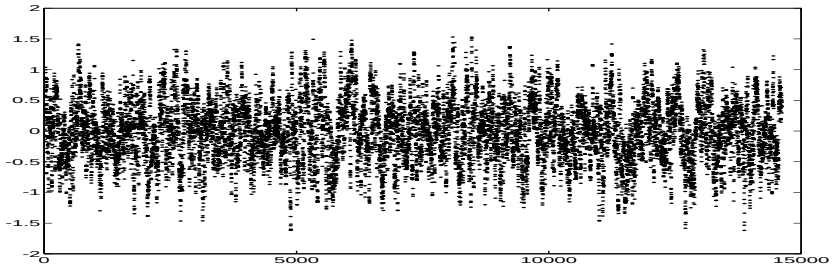
where the kernel $K$ is a density function with variance 1 and $b$ is the smoothing parameter or bandwidth. The kernel estimate of $c(t)$ is a weighted average of the observations registered in periods near $t$ and the bandwidth controls which points are considered to be in the neighbourhood of $t$. The weights for periods near $t$ are higher when the bandwidth is smaller. For large $b$ the estimate of $c(t)$ is a smooth function, while for $b$ excessively small $\hat{c}(t)$ preserves part of the variability present in the sample. Therefore, the bandwidth choice is crucial in nonparametric regression estimation. The decission about the kernel $K$ affects less in the estimation. We select the Epanechnikov kernel for its optimality properties (see Silverman, 1986). The criterion to choose the bandwidth is to reach the objectives pursued with the smoothing procedure as much as possible: the estimation of $x_t$, $\hat{x}_t = z_t - \hat{c}(t)$, should not have seasonal structure, and $\hat{c}(t)$ should be a periodic function. The first objective is achieved with bandwidths small enough, while the second requires wide bandwidths. The trade-off between both approaches allows us to overcome the difficult task of choosing the bandwidth $b$ and lead us to the value $b = 480$. This means that in the estimation of $c(t)$ we use observations separated from the period $t$ up to two months.

We estimate $c(t)$ using only observed data (no interpolation is needed at this point because the relative large size of the bandwidth). The nonparametric estimate is shown in Figure 7a, and $\hat{x}_t$ for the series with the optimal interpolation is shown in Figure 7b.

Now we fit a regular ARIMA model to $\hat{x}_t = z_t - \hat{c}(t)$. In Figure 8a the first 8000 lags of the sample ACF of $\hat{x}_t$ are shown. We observe that the seasonal autocorrelations (lags multiples of s= 2920) are non significant or they are very close to the confidence bands.




(a)




(b)

**Figure 7**: a) Nonparametric estimation of $z_t$; b) series $\hat{x}_t = z_t - \hat{c}(t)$.

Considering that the seasonal effects have been sufficiently mitigated with the smoothing procedure we identify a model for the regular structure of $\hat{x}_t$. Non stationarity is also appreciated and we take a regular difference. In Figures 8b and 8c are shown the first 100 values of the sample ACF and PACF of the differentiated series. The models we identified are listed in Table 3. The most appropriate model is a MA(1)$AR_8$(1) model, according with the criterion of parsimony in the number of parameters. Other criteria do not show strong differences between both models. Therefore, the final decision depends on which model provides better forecasts.

## 4. SHORT TERM FORECASTING

In this section we present a comparative study of the seasonal ARIMA and mixed models estimated in section 3. The objective is to select the model which generates better forecasts for future data. The interesting forecast horizon for the port authorities is two days, that is equal to 16 steps ahead predictions.

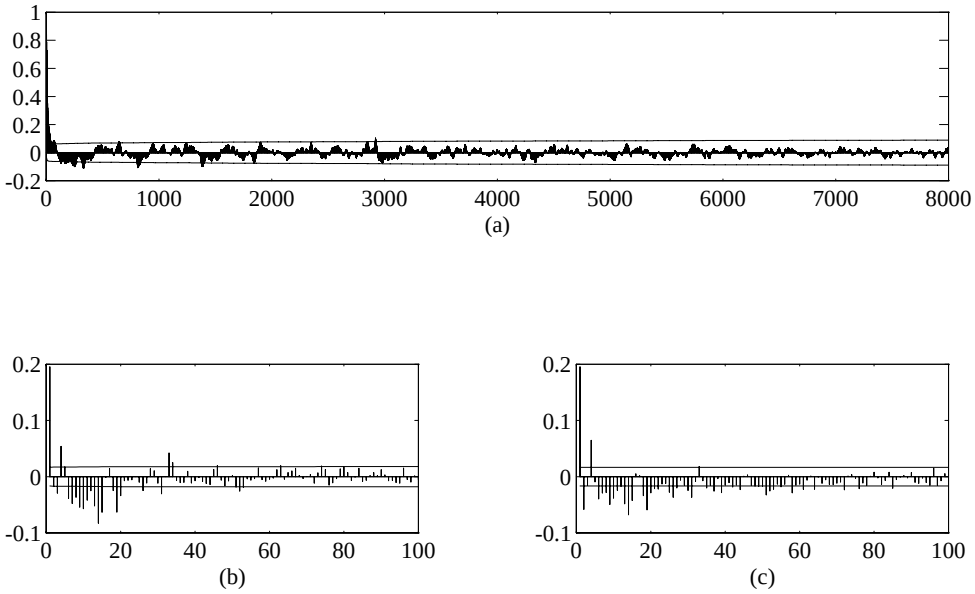



**Figure 8:** Series $\hat{x}_t = z_t - \hat{c}(t)$: a) ACF; b) ACF of the differentiated series; c) PACF of the differentiated series.

| **Model** | **Parameters** | | | | **Estimated parameters (t-value)** | | | |
|---|---|---|---|---|---|---|---|---|
| **(1)** AR(16) | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\phi_4$ | 0.194 (23.4) | -0.057 (-6.7) | -0.035 (-4.2) | 0.057 (6.8) |
| $\hat{\sigma}^2 = 0.019$, $R^2 = 0.921$ | $\phi_5$ | $\phi_6$ | $\phi_7$ | $\phi_8$ | 0.006 (0.8) | -0.038 (-4.5) | -0.030 (-3.5) | -0.023 (-2.7) |
| | $\phi_9$ | $\phi_{10}$ | $\phi_{11}$ | $\phi_{12}$ | -0.043 (-5.1) | -0.031 (-3.7) | -0.024 (-2.8) | -0.012 (-1.5) |
| | $\phi_{13}$ | $\phi_{14}$ | $\phi_{15}$ | $\phi_{16}$ | -0.037 (-4.4) | -0.059 (-7.1) | -0.043 (-5.1) | 0.005 (0.6) |
| **(2)** MA(1)AR$_8$(1) | $\theta_1$ | | | | -0.204 (-25.2) | | | |
| $\hat{\sigma}^2 = 0.020$, $R^2 = 0.919$ | $\phi_1^8$ | | | | -0.021 (-2.5) | | | |

**Table 3:** ARIMA models for $\hat{x}_t = z_t - \hat{c}(t)$ with a regular difference.

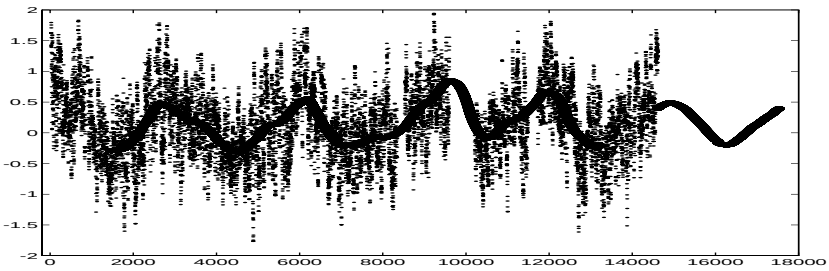



**Figure 9**: Annual cycle estimation and forecasting.

Prediction with ARIMA models is carried out in the usual way (see Box and Jenkins, 1976). However, when a mixed model is used, the forecast is the sum of the forecast with $c(t)$ and the ARIMA forecast of $x_t$. The prediction of the annual cyclic component $c(t)$ is the mean of the complete cycles estimated with nonparametric regression

$$\hat{c}_N(N+t) = \frac{1}{A}\sum_{i=1}^{A} \hat{c}(t+(i-1)s), \qquad t = 1, \ldots, s, \tag{4.3}$$

where $A$ is the number of complete years in the series. In order to avoid the effect of the ends of the series in the estimation, instead of (4.3) we suggest

$$\hat{c}_N(N+t) = \frac{1}{A-1}\sum_{i=1}^{A} \hat{c}(t+(i-1)s) I_{[a_1,a_2]}(t+(i-1)s), \qquad t = 1, \ldots, s, \tag{4.4}$$

where $a_1 = (1 + s/2)$ and $a_2 = ((A-1)s + s/2)$. The equation (4.4) is the mean cycle of the four complete cycles estimated from the second semester of 1986 to the end of the first half of 1990. The estimate $\hat{c}(t)$ for the observed years and the forecast $\hat{c}_N(N+t)$ for the first year outside of the sample are shown in Figure 9.

We predict for 20 days with the four models identified and estimated in section 3. Every two days we obtain 16 steps ahead predictions. The forecast origin is at 9:00 p.m. on 12-31-1990, the last data used in the model estimation. The model comparison is based on the forecast mean square error: an average of the distance between the real values and the forecasts. We know the actual data registered during the twenty days in which the forecasts are done.

The forecast mean square errors are specified in Table 4 for each of the four models when the forecast horizons go from 1 to 16 steps ahead. The last line in the table shows

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

**Table 4:** Forecast mean square error with seasonal ARIMA and mixed models ($s = 2929$).

the mean values of the mean square errors. We observe that except for the one step ahead prediction, the mixed models always overcome the pure ARIMA models. We also appreciate that to include the autoregressive polynomial of order 16 in the regular part provides the best forecasts with mixed models. Therefore, the most appropriate model in order to predict with a two days forecast horizon is the mixed model with linear component $AR(16)$:

$$(1 - 0.194B + 0.057B^2 + \ldots + 0.035B^{16})(1 - B)(z_t - \hat{c}(t)) = a_t.$$

## 5. CONCLUSIONS

The finally proposed model to short term forecast the significant wave height is a mixed model in which we fit an AR(16) to the linear part, free from the cyclic behavior of the series. We have seen that this is the model with minimum forecast mean square error in the 20 days where several models are compared. Another advantage of this model is that the linear part not include seasonal differences. For this reason, in forecasting tasks, this model demands less computer time than models including seasonal differences do.

To improve the system performance, the model should be updated periodically. We propose to estimate the cycle $c(t)$ each half year and to revise the parameter estimation of the AR(16) polynomial every month.

## ACKNOWLEDGMENTS

The authors thank to Daniel Peña and Juan Romo their valuable comments and to Obdulio Gómez and José Carlos Nieto to facilitate the series. Authors are also grateful to the Department of Statistics and Econometrics at Universidad Carlos III de Madrid to provide them with computing facilities. This work is partially financed by the Ente Público Puertos del Estado and by the project PB93–0232 from the DGYCIT.

## REFERENCES

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

Brubacher, S.R. and 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,* Tokyo: University of Tokyo Press.

Härdle, W. (1990). *Applied Non-parametric Regression,* London: Oxford University Press.

Maravall, A. and Peña, D. (1997). 'Missing observations and additive outliers in time series models', *Advances in Statistical Analysis and Statistical Computing,* JAI Press (in press).

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

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