ABOUT A PROMPT STRATEGY FOR ESTIMATING MISSING DATA IN LONG TIME SERIES

por

Fabio H. Nieto[1] & Fernando Ruiz[2]

Resumen

Fabio H. Nieto & Fernando Ruiz: About a prompt strategy for estimating missing data in long time series. Rev. Acad. Colomb. Cienc. **26**(100): 411-418, 2002. ISSN 0370-3908.

Se propone una rápida estrategia práctica para estimar datos faltantes en series temporales que obedecen a modelos ARIMA de orden bajo y cuya longitud es mayor que la que soportan programas de cómputo estadístico. La metodología propuesta se basa en la idea de identificar el modelo para la serie a partir de sus subseries. Para obtener estas subseries, un número mínimo de datos después de una observación faltante se deduce para lograr una estabilización numérica de su predicción recurrente.

Palabras clave: Modelos ARIMA, observaciones faltantes, identificación de modelos.

Abstract

A quick practical strategy is proposed for estimating missing data in time series that obey low-order ARIMA models and whose length is greater than that supported by current statistical computer programs. The proposed methodology is based on the idea of identifying the series model from subseries of it. For obtaining the subseries, a minimal number of data after a missing observation is deducted for achieving numerical stabilization in its recursive prediction.

Key words: ARIMA models, Missing observations, Model identification.

[1]Departamento de Estadística, Universidad Nacional de Colombia, Bogotá. e-mail: fnietomatematicas.unal.edu.co
[2]IDEAM, Bogotá, Colombia.

1. Introduction

Sometimes, we are faced with the problem of estimating missing observations in *very long* time series. This is the case for example of the daily flow of a river in which we can have more than 10,000 observed data (more than 30 years) with *many* missing observations. Efficient statistical estimation of these missing data is important because, in doing so, we can get official complete data bases and then users in the agricultural, environmental, and natural resources fields can fit statistical models with several related variables without losing information in these variables.

At present, we do not have a commercial statistical computer program for solving numerically the problem quoted above using, for example, Gomez and Maravall's [6] efficient methodology. This is due basically to memory restrictions for handling simultaneously (1) model identification in the presence of missing data, (2) exact maximum-likelihood estimation of the model parameters taking into account the missing observations, and (3) interpolation of the missing data using additive-outlier-model based approaches. One of the most appealing programs for doing this task with time series data is TRAMO (Gomez and Maravall, [7]) but the program only permits a length of at most 2000 data. In terms of daily data this represents about 5.5 years and, for instance, this cannot be enough information in hydrological contexts. In what follows, the word *long* will be used in the sense that the time series length is greater than the maximum permitted by all the current programs for analyzing time series *with missing observations.*

Of course, this computing problem does not exist if one has at hand a powerful computer where whatever computer program that estimates missing data can run with long time series without memory restrictions. Thus, at least two ways for solving this problem are: (1) to find an appropiate machine and develop the actual programs for it, or (2) to develop a prompt reasonable solution based on the actual programs. At agencies or institutions that everyday have to deal with data analysis and need quick strategies, the second alternative is appealing.

Recently, Delicado and Justel [4] solved the aforementioned long-time-series problem using essentially (1) a "natural" interpolation approach for completing the series in order to identify and estimate an ARIMA model and (2) the dual autocorrelation function of this model for interpolating the time series. Some eventual drawbacks could appear in Delicado and Justel's work. Since the "natural" interpolation consists in using *means* as preliminary estimated values for missing data, the identified model could be not appropriate if each subseries of the true one is nonstationary. As a consequence, one could obtain inefficient model parameter estimates using the "complete" series in place of the uncompleted one. For this last case, the appropriate procedure is that of Gomez and Maravall [6]. Since the dual autocorrelation function needs to be truncated when the model contains an MA factor, the interpolations obtained with this tool could be suboptimal as indicated by Nieto [8], especially if one has missing observations at the ends of the series. The appropriate way is using smoothing as indicated by Gomez and Maravall [6]. Obviously, with relatively few missing data, no missing observations at the ends, and a computer package that handles *long* time series, Delicado and Justel's approach is appealing.

In this paper, we propose a *practical* strategy for an adequate use of Gomez and Maravall's [6] methodology for estimating missing data in long time series which obey low-order nonseasonal ARIMA models, as is the case for hydrological/meteorological daily data. Essentially, the strategy consists in splitting the long series into *subseries* for (1) identifying the model, (2) estimating its unknown parameters, and (3) interpolating the missing observations. This is an idea that was also used by Politis *et al.* [11] in the solution of an econometric problem and by Tong and Lim [12] for identifying a threshold autoregressive model.

In Section 2 we present some basic theoretical and numerical results that are the basis for the practical methodology developed in Section 3. Section 4 includes simulated and empirical examples for illustrating the proposed practical approach. Finally, Section 5 concludes.

2. Some theoretical and numerical results

2.1 Theoretical background. Let us assume that the stochastic process $\{Z_t\}$ follows the ARIMA model

$$\phi(B)\delta(B)Z_t = \theta(B)a_t \ ,$$

where $\phi(B)$, $\delta(B)$, and $\theta(B)$ are finite polynomials in the lag operator B, and $\{a_t\}$ is a zero-mean Gaussian white noise process with variance σ^2. $\phi(B)$ contains the stationary roots, and $\delta(B) = 1 - \delta_1 B - ... - \delta_d B^d$ contains the possible unit roots so that $\{W_t\} = \{\delta(B)Z_t\}$ is a stationary process.

Suppose now that $\{Z_t\}$ has been observed from $t = 1$ up to $t = T$ and that there are k missing data in the sample at time points $m(1), ..., m(k)$ with $d < m(1) < ... < m(k)$. Let us label the time points after $t = d$ where there are observed data as $t(1), ..., t(r) = T$ with $r = T - d - k$. Let $P_j Z_{m(l)}$, $l = 1, ..., k$, be the best linear predictor (BLP) of $Z_{m(l)}$, in the sense of minimum mean square error (MMSE), given the variables Z_1,...,Z_d,$Z_{t(1)}$,...,$Z_{t(j)}$, where $j = 1, ..., r$ [see Catlin [3] or Brockwell and Davis [2] for a formal definition of BLP]. Gomez and Maravall [6] and Nieto [9] have obtained *exact* recursive formulas for computing $P_j Z_{m(l)}$ which are based only on the *finite* observed sample. We shall also use this finite-sample context here.

In the Appendix, we show that the sequence $\{P_j Z_{m(l)}\}$ is mean-square convergent where l is fixed and j varies in such a way that $t(j) > m(k) + d$. Consequently, $\{P_j Z_{m(l)} - P_{j-1} Z_{m(l)}\}$ converges to 0 in mean square as $j \to \infty$. This is an important result for setting numerical stopping criteria in practice. That is, for establishing a minimal number of data after $m(k) + d$ for which $\{P_j Z_{m(l)}\}$ stabilizes.

The rate of convergence towards zero of $\{P_j Z_{m(l)} - P_{j-1} Z_{m(l)}\}$ depends on the autocorrelation structure of $\{W_t\}$ (see Nieto, [9]) or, equivalently, on the ARMA parameters of $\{W_t\}$. Since it is very difficult to determine this dependence by analytical means, we concentrate only on nonseasonal invertible ARIMA(p, d, q) models with $p, q \leq 1$ and $d \leq 2$, to analyze this kind of problem and we shall proceed via simulation. For future reference, these ARIMAs will be called *low-order* models.

Remark. The difficulty of finding the dependence of the predicted missing value on the autocorrelation structure of $\{W_t\}$ is due to the finiteness characteristic of the time series. In the case of infinite samples, an ideal theoretical situation, it is known that this dependence is determined by the Dual Autocorrelation Function of the process $\{W_t\}$ [Peña and Maravall [10])].

The reasonableness of using low-order models comes from the facts that (1) it is very difficult to have the usual assumptions behind seasonal models (Brockwell and Davis [2]) fulfilled by daily data, especially hydrological data, (2) practice indicates that, for example, daily hydrological or meteorological series are well described by ARMA(p, q) models where p, q are not very large, (3) practice also shows that for interpretational purposes, d must be less or equal than 2 , and (4) the simulation study carried out in the next subsection can be easily imitated for more general models, if necessary.

2.2 Simulation results. The interest now is in determining a *minimum* value of j for which the sequence $\{P_j Z_{m(l)}\}$ stabilizes quickly, independent of the autocorrelation structure of the process $\{W_t\}$. That is, given $\varepsilon > 0$ small enough, determine j_ε such that $|P_j Z_{m(l)} - P_{j-1} Z_{m(l)}| < \varepsilon$ for all $j > j_\varepsilon$. Without loss of generality, we consider only the case $k = 1$. This task will be done via simulation and, with that end, we design the experiment in the following way:

Following the survey-sampling idea of obtaining a maximum sample size for estimating a population proportion given a pre-established margin of error, we put $d = 2$ and $m(1) = 3$ (the extreme case). Taking into the account the dependence between the missing-value predictor and the autocorrelation structure of $\{W_t\}$, we consider the following ARMA parameter values: (i) if $p = 1$ and $q = 0$, we take $\phi = 0.1, 0.2, 0.4, 0.6, 0.8, 0.9$; (ii) if $p = 0$ and $q = 1$, we put $\theta = 0.2, 0.4, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99$; and (iii) for $p = q = 1$, we use pairs (ϕ, θ) with ϕ in the set of (i) and θ in the set of (ii).

Of course, a more refined partition of the parameters space could be considered, but, as was indicated by the simulation results, this is not necessary.

Since in computing $P_j Z_{m(1)}$ the white noise variance σ^2 is dropped out and in calculating its MSE it becomes a scale factor (Nieto, [9]), we set $\sigma^2 = 1.0$ in the simulation process. The basic results for the MA(1) and ARMA(1,1) models are presented in Tables 1, 2, and 3 where the entries are the minimal values j_ε for corresponding values of ε and ϕ or θ. In the AR(1) case, we found that if $\varepsilon = 10^{-6}$, $j_\varepsilon = 3$ for all ϕ in the set considered, and for $\varepsilon \in \{10^{-3}, 10^{-4}, 10^{-5}\}$, we obtained $j_\varepsilon = 2$ for all ϕ. All the numerical results were obtained using the Fixed Point Smoother algorithm developed by Gomez and Maravall [6], which is theoretically equivalent to Nieto's [9] procedure.

The interpretation of the results for the MA(1) and ARMA(1, 1) models is the following. In Table 1, if we fix ε, j_ε increases when θ increases and large values are obtained for θ very close to 1. Similarly, if we fix θ and let ε decrease, j_ε increases, as expected, also with large values when θ is close to 1. Although we only include values of $\theta \geq 0.5$, the previous observation is valid for all θ in the set considered.

For the ARMA(1,1) model, Table 2, we found that relatively few data are needed for achieving numerical convergence in the sense that $9 \leq j_\varepsilon \leq 17$ for ε in the

set $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ and for (ϕ, θ) in the range considered.

The dependence of j_ε on ϕ is stronger than on θ and this parameter has influence on j_ε when this value is compared to the corresponding one for the AR(1) model. We note that, for space limitations, all the values of θ were not included; however, the pattern observed in the table is the same that we found for all θ considered in the simulation. Additionally, we can see that curiously there are "discontinuities" in j_ε at points $(0.1, 0.4)$, $(0.6, 0.8)$, and $(0.9, 0.4)$ where its values are respectively 3, 1, and 1. In Table 3 we summarize the results presented in Table 2 where each entry is the maximum of j_ε over θ fixing ϕ and ε.

In general terms: (i) If the true low-order model for the process $\{W_t\}$ contains an AR factor, we need a number of data between 9 and 17, after point $m(k) + d$, to get a precision greater or equal than 10^{-6}. (ii) In a drastically contrary way, if the model is MA(1) with θ very close to 1, we need about 868 observations after $m(k) + d$ to get a precision of about 10^{-6} in the missing data predictions. For a reasonable $\theta \leq 0.9$ and tolerance $\geq 10^{-6}$, this figure decreases severely to about 104 data. Of course, true values of θ near to 1 can induce invertibility problems in practice or possible overdifferentiation of an observed series.

The practically important point deduced from these results is this: in the absence of any kind of knowledge about the mathematical form of the model, except that it is low-order with $\theta \leq 0.9$, we should take the number of data j_ε between 39 and 104, after the point $m(k) + d$, for obtaining numerical convergence of the sequence of missing data predictions with a precision ε between 10^{-6} and 10^{-3}. For example, for a precision of about 10^{-4}, 60 data is a reasonable value.

3. The proposed practical methodology

Let n be the maximum number of data that a statistical package for estimating missing data in a time series supports. Let $\{z_1, ..., z_T\}$, $T > n$, be the observed time series from the stochastic process $\{Z_t\}$ which obeys the low-order ARIMA model given in Section 2.

Now, we divide the given time series into the subseries

$$\{z_1, ..., z_{T_1}\}, \{z_{T_1+1}, ..., z_{T_2}\}, ..., \{z_{T_{K-1}+1}, ..., z_{T_K}\} ,$$

where $1 < T_1 < ... < T_{K-1} < T_K = T$. The cut points $T_1, ..., T_{K-1}$ are chosen in such a way that (i) $T_j - T_{j-1} \leq n$ for all $j = 1, ..., K$, with $T_0 = 1$, and (ii) the numerical result obtained in the last section is satisfied in the sense that if there exists $j_i \in \{1, 2, ..., K\}$ for some $i \in \{1, 2, ..., k\}$ such that $m(i) < T_{j_i}$, then $T_{j_i} \geq m(i) + d + j_\varepsilon$ for some pre-established j_ε. The purpose is to identify the ARIMA model from these K subseries, taking into account that, in practice, different subseries can lead to different identified models.

Table 1. Simulation results for the MA(1) model; the values of j_ε in the entries

	θ					
ε	0.5	0.8	0.9	0.95	0.99	Marginal maximum
10^{-3}	8	21	39	66	183	183
10^{-4}	11	31	60	111	409	409
10^{-5}	15	42	82	156	638	638
10^{-6}	18	52	104	201	868	868

Table 2. Results of the ARMA(1,1) model, with values of j_ε in the entries

		ε			
ϕ	θ	10^{-3}	10^{-4}	10^{-5}	10^{-6}
0.1	0.2	9	11	15	17
	0.4	3	11	15	17
	0.6	9	11	15	17
	0.9	9	11	14	17
	0.99	9	11	14	17
0.2	0.4	7	11	13	16
	0.6	8	11	13	16
	0.9	8	11	13	16
	0.99	8	11	13	16
0.4	0.2	7	9	11	14
	0.6	7	9	12	14
	0.8	7	9	12	14
	0.9	7	9	12	14
	0.99	7	9	12	14
0.6	0.2	7	9	11	13
	0.7	7	9	11	13
	0.8	1	9	11	13
	0.9	6	8	11	13
	0.99	6	8	11	13
0.8	0.2	7	9	11	12
	0.4	7	9	10	12
	0.6	6	8	10	12
	0.9	6	8	10	12
	0.99	6	8	10	12
0.9	0.2	7	9	11	13
	0.4	1	9	10	12
	0.6	6	8	10	12
	0.8	6	8	10	12
	0.99	6	8	10	11
Marginal maximum		9	11	15	17

Since each subseries contains *information*, in the Kullback-Leibler sense [see Brockwell and Davis [2] for the meaning of this concept], about the true underlying probability mechanism, it seems reasonable to hope that one can gain information about a global statistic by evaluating it on all the subseries (or "subsamples") and then combining optimally the marginal statistics to obtain the global one. This is the philosophical principle under which Politis *et al.* developed their paper and the combining-forecasts practice is performed. A similar idea was also used by Tong and Lim [12] for identifying threshold autoregressive models. Nevertheless, the theoretical formality of our approach will be investigated in future work.

Table 3. ARMA(1,1)model summary; values of j_ε in the entries

	ϕ						
ε	0.1	0.2	0.4	0.6	0.8	0.9	Marginal maximum
10^{-3}	9	8	7	7	7	7	9
10^{-4}	11	11	9	9	9	9	11
10^{-5}	15	13	12	11	11	11	15
10^{-6}	17	16	14	13	13	13	17

The strategy is the following:

Stage I. *Model Identification.*

We firstly examine the assumption that $\{Z_t\}$ obeys a low-order ARIMA process proceeding as follows:

A. With an automatic identification procedure applied to each subseries we obtain its respective model and its Bayesian Information Criteria (BIC).

B. If low-order models are the most frequent ones, then we accept our assumption and continue with step C. Otherwise we stop the procedure because our method is not applicable.

C. Among the low-order models, select the most frequent one as the appropriate model for the whole process $\{Z_t\}$. If there are two or more models that have the same high frequency of occurrence, choose among them the model with the lowest number of missing data. If this number of minimal missing data is shared by several of these models, choose the model with the minimum BIC.

Stage II. *Interpolation of the series.*

D. Interpolate each of the K subseries with the model identified in STAGE I, reestimating firstly the model parameters in each case. Now, if there exists a subseries for which was impossible to choose a cut point with the condition indicated above, the last missing data in that subseries must be reestimated using step E below.

E. From the whole series $\{z_1, ..., z_T\}$, obtain n-data subseries centered at the points T_{j_i} for which the corresponding is are such that $T_{j_i} < m(i) + d + j_\varepsilon$. Within these subseries make the interpolation as in D.

Remark. Step E is justified by the following. Prediction of a missing observation before some T_j, for which the numerical-convergence criterion is not satisfied, is suboptimal because its predictor is not using the minimal number of data deduced for such purpose. In using Step E, we are including additional data in the time period $[T_j + 1, T_{j+1}]$, although it can happen that they are not still enough for satisfying the minimal requirement for convergence. Nevertheless, this last estimate is more efficient than that obtained in Step D.

4. Some examples

Example 1. We simulate an ARIMA(1,1,1) model with parameters $\phi = 0.5$, $\theta = 0.8$, $\sigma^2 = 100$, and without constant. Program RATS (Doan, [5]) was used with seed 14600 and the initial condition z_1 was generated with a N(0,1) distribution. To obtain z_2 we used Bell's [1, Theorem 1]. The length of the simulated series was $T = 14600$, that is, approximately 40 years in terms of daily data.

With $n = 600$ and $\varepsilon = 10^{-4}$ ($j_\varepsilon = 60$) we obtained 24 subseries from the simulated series and the results obtained from STAGE I, using TRAMO, were the following: (i) transformation of data was not identified, (ii) except in a subseries, we do not identified constant for the model, and (iii) the simulated model was identified with a frequency of 66.7%. All the identified models were low-order.

To illustrate the adequateness of the methodology in the missing data case, we take out some data in 13 subseries. The minimum number of missing data considered was 5 and the maximum 17. The identified model was again the simulated one although less frequent than in

the complete data case. Again, all the ARIMA identified models were low-order.

Example 2. Here we consider a real data application. The series consists of 7639 daily observations of the Magdalena river flow in Colombia made at El Contento station. This type of data is assembled by IDEAM, the Colombian official agency for hydrological and meteorological studies and can be obtained from the authors upon request. The series contains 1439 missing data (about 19% of the sample size) which are spread on all the series. Actually, it is known in the literature that hydrological/meteorological time series are well described by stationary nonlinear models (Tong, [13]). However, in this example we are going to transform the data to get some degree of linearity in order to use the proposed methodology.

Using the automatic model identification procedure implemented in TRAMO, a log transformation and a differentiation were suggested by the data. Here, we must be cautious in interpreting these transformations. What they are meaning is that the relative changes of the series are approximately linear; hence, this is the objective series to be analyzed. Since the data are originally stationary, we obtain an overdifferentiation of the time series, which, in terms of parameter estimation, is not a serious problem when the process is ergodic. It is not the case that this differentiation indicates a unit root in the original data, which would not make sense in a hydrological flow time series.

Table 4. Identified models for the subseries in the real–life example .

Subseries No.	Size	Model	BIC	No. of miss.data
1	600	(1,0)	9.13	0
2	600	(1,2)	9.03	89
3	533	(1,0)	9.30	59
4	600	(1,0)	9.50	90
5	527	(1,1)	9.15	90
6	578	(1,1)	9.70	152
7	488	(1,0)	10.20	242
8	600	(4,0)	10.56	301
9	526	(4,0)	9.82	91
10	600	(1,1)	9.50	87
11	600	(0,1)	−5.89	0
12	600	(0,1)	9.72	121
13	583	(1,0)	8.92	93
14	204	(1,1)	9.02	24

Table 4 presents the basic results about the identified model for the whole transformed series. We can observe that 86% of the models are low-order and that, among them, the ARMA(1,0) model was the most frequent one (36% of the times). Because of subsampling variability, two no low-order models (14%) were identified. Based on these results, our decision is not to reject the assumption of a low-order ARMA model for the relative changes of the time series and to pick up the ARMA(1,0) as the appropriate model for it. With this model we interpolate the missing values in each subseries using the additive-outlier approach, where possible, or the skipping approach (fixed point smother) both implemented in TRAMO.

5. Conclusions

We have developed a quick practical approach for estimating missing data in time series which (i) have a length greater than that supported by current statistical computer programs and (ii) are supposed to follow an ARIMA(p, d, q) model with $p, q \leq 1$ and $d \leq 2$; as is the case for example of daily hydrological data.

Using a simulation study, we have found that, independent of the correlation structure, a number of data between 39 and 104 after the last missing observation is adequate for achieving (numerical) convergence in the sequence of missing data predictors with a precision between 10^{-6} and 10^{-3}. This number of data increase if the desired precision increase or the moving average parameter is close to 1. The study can be imitated for solving the same problem in higher order models.

The proposed strategy identifies the model for the whole series from identified models for subseries of it, which is necessary for computing the missing data estimates. For splitting the series, we use the numerical-convergence result mentioned above and as the global model, the most frequent among the low-order identified models. In future research, this identification-model procedure will be formalized by means of a statistical test for the null hypothesis that a given series obeys a *low-order* ARIMA model.

Acknowledgments

This is a substantially improved version of a consulting job about the topic, that the first author carried out for IDEAM under contract 037/98. The authors gratefully thanks Emperatriz Español at IDEAM and Fernando Martinez and Javier Forero (students of the undergraduate Statistics program at Universidad Nacional de Colombia) for the benefit of discussions on the topic. Additionally, they also acknowledge Professor Jairo Charris for a useful suggestion about the mathematical result in the Appendix.

Appendix

Proposition. *The sequence of predictors* $\{P_j Z_{m(l)}\}$ *converges in mean-square to the BLP of* $Z_{m(l)}$ *based on* $Z_1,...,Z_d,Z_{t(1)}, Z_{t(2)},...,Z_{m(k)+1},\cdots$.

Proof. Let

$$S = \bar{sp}\{Z_1, ..., Z_d, Z_{t(1)}, Z_{t(2)}, ..., Z_{m(k)+1}, ...\} ,$$

where $\bar{sp}$ denotes the closed span of a set of random variables. See Catlin [3] or Brockwell and Davis [2] for this definition. Let $PZ_{m(l)}$ be the orthogonal projection of $Z_{m(l)}$ onto S. Following those authors, $PZ_{m(l)}$ is the BLP of $Z_{m(l)}$ based on $\{Z_1, ..., Z_d, Z_{t(1)}, Z_{t(2)}, ..., Z_{m(k)+1}, ...\}$, which always exist and is unique. To show that $\{P_j Z_{m(l)}\}$ converges in mean square to $PZ_{m(l)}$ as $j \to \infty$ we proceed as follows:

Let $\{Y_1, ..., Y_d, Y_{t(1)}, Y_{t(2)}, ..., Y_{m(k)+1}, ...\}$ be a orthogonalization of the set

$$\{Z_1, ..., Z_d, Z_{t(1)}, Z_{t(2)}, ..., Z_{m(k)+1}, ...\} ,$$

then

$$PZ_{m(l)} = \sum_{i \in I} E(Y_i Z_{m(l)}) Y_i$$

and

$$P_j Z_{m(l)} = \sum_{i \in I_j} E(Y_i Z_{m(l)}) Y_i ,$$

where $I = \{1, ..., d, t(1), t(2), ..., m(k)+1, ...\}$ and

$$I_j = \{1, ..., d, t(1), t(2), ..., m(k)+1, ..., t(j)\} .$$

Hence

$$E\{(PZ_{m(l)} - P_j Z_{m(l)})^2\} = \sum_{i \in I'_j} [E(Y_i Z_{m(l)})]^2 \quad \text{(A1)}$$

where I'_j is the complement set of I_j. Now, since $\sum_{i \in I} [E(Y_i Z_{m(l)})]^2 < \infty$, then the right member in (A1) converges to 0 as $j \to \infty$. This ends the proof.

References

[1] **Bell, W. R.**, *Signal extraction for nonstationary time series.* The Annals of Statistics, **120**(1984), 646-664.

[2] **Brockwell, P. J. & Davis, R. A.**, *Time Series: Theory and Methods.* Springer Verlag, New York, 1991.

[3] **Catlin, D. E.**, *Estimation, Control, and the Discrete Kalman Filter.* Springer Verlag, New York, 1989.

[4] **Delicado, P. & Justel, A.**, *Forecasting with missing data: Application to coastal wave heights.* Journal of Forecasting, **18** (1999), 285-289.

[5] Doan, T. *Regression Analysis of Time Series (RATS)*, Estima, Evanston, IL, 1992.

[6] **Gomez, V. & Maravall, A.**, *Estimation, prediction, and interpolation for nonstationary series with the Kalman filter.* Journal of the American Statistical Association, **89** (1994), 611–624.

[7] **Gomez, V. & Maravall, A.**, *PROGRAMS TRAMO AND SEATS: Instructions for the user*, Banco de España–Servicio de Estudios, Documento de Trabajo No. 9628, Madrid, 1996.

[8] **Nieto, F. H.**, *Una nota sobre la estimacion de datos faltantes en una serie temporal, usando la funcion de autocorrelacion dual.* Estadistica **46** (1994), 85–103.

[9] **Nieto, F. H.**, *A note on interpolation of ARIMA processes.* Communications in Statistics, Theory and Methods **26**, No. 10,(1997).

[10] **Peña, D. & Maravall, A.**, *Interpolation, outliers, and inverse autocorrelations.* Communications in Statistics: Theory and Methods **20** (1991), 3175–3186.

[11] **Politis, D.Ñ., Romano, J. P., & Wolf, M.**, *Subsampling for heteroskedastic time series.* Journal of Econometrics **81** (1997), 281–317.

[12] **Tong, H. & Lim, K. S.**, *Threshold autoregression, limit cycles, and cyclical data.* Journal of the Royal Statistical Society, Series B, **42** (1980), 245–292.

[13] **Tong, H.**, *Nonlinear Time Series*, Oxford University Press, Oxford, 1990.