TECHNOMETRICS, AUGUST 1995, VOL. 37, NO. 3

Predicting Using Box–Jenkins, Nonparametric, and Bootstrap Techniques

I. GARCÍA-JURADO, W. GONZÁLEZ-MANTEIGA, J. M. PRADA-SÁNCHEZ, and M. FEBRERO-BANDE

Department of Statistics and O.R.
University of Santiago de Compostela
Santiago de Compostela
Spain

R. CAO

Department of Mathematics
University of La Coruña
La Coruña
Spain

In this article, a new semiparametric prediction system is presented for time series. The prediction method incorporated to the system consists of a nonparametric part that estimates the trend, a Box–Jenkins prediction of the residual series, and some bootstrap methodology to construct prediction intervals. Consistency of the estimators proposed for the autoregression function and the parameters in the Box–Jenkins model and the validity of a new bootstrap resampling plan adapted to autoregressive integrated models are proved. The Monte Carlo simulation study, as well as the applications to real data (carried out with the automated system, incorporating the method, developed for predicting concentration levels in the surroundings of a Spanish power station), show that this method outperforms other standard competitors.

KEY WORDS: Concentration levels; Kernel estimation; Semiparametric; Time series.

1. INTRODUCTION

Let (Z_l, Y_l), $l = 0, \pm 1, \pm 2, \ldots$, be a strictly stationary series, where Z_l is an r-dimensional series and Y_l is a one-dimensional response series. We want to estimate $\varphi(z_l^0) = \varphi(F(. \mid Z_l = z_l^0))$, where $F(. \mid Z_l = z_l^0)$ is the conditional distribution of Y_l given $Z_l = z_l^0$, using a series $\{(Z_i, Y_i)\}$ of length n. Frequently, φ is the mean or median functional. In particular, when $Y_l = X_{l+k}$, $k \geq 1$ and $Z_l = (X_l, \ldots, X_{l-r+1})$, where X_l is a stationary series, we are estimating the autoregression function of order k,

$$\varphi(x_1^0, \ldots, x_r^0) = E(X_{l+k} \mid (X_l, \ldots, X_{l-r+1}) = (x_1^0, \ldots, x_r^0)), \quad (1)$$

using the sample $\{X_{t-m+1}, \ldots, X_t\}$ of size m.

In the literature of time series, there are essentially two approaches. The Box–Jenkins methodology approximates an autoregression function (1) using a linear combination of variables by minimizing the mean squared prediction error. For an (autoregressive) AR(p) model, $X_l = \psi_0 + \psi_1 X_{l-1} + \cdots + \psi_p X_{l-p} + a_l$, where $\{a_l\}$ is white noise, r equals p, and the autoregression function of order k is given by $\varphi(x_1^0, \ldots, x_p^0) = \psi_0^{(k)} + \psi_1^{(k)} x_1^0 + \cdots + \psi_p^{(k)} x_p^0$, where the coefficients $(\psi_0^{(k)}, \ldots, \psi_p^{(k)})$ are recursively derived in k steps from the one-lag coefficients $(\psi_0, \ldots, \psi_p)$ using the model equation (for more details, see Wei 1990, p. 89). For an (autoregressive moving average) ARMA(q, s) model given by $\psi(B)X_l = \theta(B)a_l$, where $\psi(B) = (1 - \psi_1 B - \cdots - \psi_q B^q)$ and $\theta(B) = (1 - \theta_1 B - \cdots - \theta_s B^s)$, there does not exist linear autoregression function of order k. Yet, the optimal linear predictor $\hat{X}_{t+k} = \psi_{0m}^{(k)} + \psi_{1m}^{(k)} X_t + \cdots + \psi_{mm}^{(k)} X_{t-m+1}$ is constructed using the Durbin–Levinson algorithm (Brockwell and Davis 1991, chap. 5) with an initial sample of size m. In both cases the ψ coefficients must be estimated using the observed sample (note that the number of coefficients to be estimated is not the same in the AR and the ARMA cases).

The second approach is based on nonparametric procedures. The function $\varphi(z_l^0) = E(Y_l \mid Z_l = z_l^0)$ is estimated directly without making any parametric assumption on it. In general, given a sample $\{(Z_i, Y_i)\}_{i=1}^n$, the estimator is given by

$$\hat{\varphi}_n(z_l^0) = \sum_{i=1}^n W_{ni}(z_l^0, (Z_1, Y_1), \ldots, (Z_n, Y_n)) Y_i, \quad (2)$$

where $\{W_{ni}\}$ can be a sequence of kernel weights, k_n–nearest neighbors, and so forth. This approach to prediction is newer. The following are outstanding works on the topic:

1. Yakowitz (1985). In this article, $Y_t = X_{t+1}$ is predicted from $Z_t = X_t$ using a sample $\{X_1, \ldots, X_m\}$ in a Markovian stationary model. The function $\varphi(x) = E(X_{l+1} \mid X_l = x)$ is estimated from (2) using kernel weights

$$W_{ni}(x, (X_1, X_2), \ldots, (X_n, X_{n+1})) = \frac{K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)}, \quad i = 1, \ldots, n - m - 1, \quad (3)$$

where K is the kernel function and h_n is the bandwidth parameter. Yakowitz presented in this work the prediction $\widehat{\varphi}_n(X_t)$ as a good alternative to ARMA models when predicting river floods.

2. Yakowitz (1987). In this article, $Y_t = X_{t+1}$ is predicted from $Z_t = (X_t, X_{t-1}, \ldots, X_{t-p_1+1}, U_t, \ldots, U_{t-p_2+1})$, where $p_1 + p_2 = p$. X_l is a stationary series, generally Markovian, U_l is an exogenous stationary series, and the sequence of weights is of the k_n–nearest neighbors type,

$$W_{ni}(z_l^0, (Z_1, Y_1), \ldots, (Z_n, Y_n)) = \frac{\left(1_{\{\|z_l^0 - Z_i\| \le R(n)\}}\right)}{k_n}, \quad i = 1, \ldots, n, \quad (4)$$

constructed using a sample of size m of X_l and the corresponding exogenous U_l with $p_2 \le p_1$ and $n = m - p_1$.

Our interest in prediction arises from a study of the contamination produced by SO_2 near a Spanish power station. In Spain, the mean of the SO_2 concentrations (recorded every five minutes), observed during a two-hour period, cannot exceed certain permitted values. To maintain control over SO_2 concentrations, the power stations must take action with a lead time of about a half hour when the observed concentrations reach disturbingly high levels. At each time t we receive a new observation X_t of SO_2 concentration and must predict X_{t+6} knowing $X_t, X_{t-1}, \ldots$. From experience, we know that the observed time series has special features—for example, long periods of stationarity, sudden and sharp explosions, or *episodes* corresponding to uncontrolled increasings of SO_2 concentrations. The lack of stationarity of the series during the episodes suggests that Box–Jenkins methodology does not forecast well during an episode. (Recall that the main goal of forecasting is to prevent high SO_2 concentrations; this can happen only when an episode arises.) Figure 3, Section 4, shows the poor behavior of Box–Jenkins methodology during an episode recorded in the surroundings of the As Pontes power station in 1990 (see Sec. 4 for more details).

In this article, we present an alternative to Box–Jenkins methodology by generalizing the preceding nonparametric model to a semiparametric model. Consider the model

$$Y_l = \varphi(Z_l) + e_l, \quad (5)$$

where e_l has an ARMA(q, s) structure independent of Z_l; we will focus on predicting Y_t after observing the series Y_l up to time $(t - k)$ and Z_l up to time t. In particular, using the sample $(Z_{t-n+1-k}, Y_{t-n+1-k}), \ldots, (Z_{t-k}, Y_{t-k})$ of size n, the prediction $\widehat{Y}_t$ of T_t is defined by

$$\widehat{\varphi}_n(Z_t) + \dot{e}_t, \quad (6)$$

where $\widehat{\varphi}_n$ is the nonparametric estimate given by (2) with weights of the type (3), (4), and so forth and $\dot{e}_t$ is the Box–Jenkins prediction, k instants ahead, constructed from the estimated ARMA component of the series $\widehat{e}_t = Y_t - \widehat{\varphi}_n(Z_t)$. Our interest in this semiparametric model comes from our experience with real data on SO_2 concentrations because the residuals from the nonparametric prediction stage are generally not white noise. This suggests that the nonparametric predictions might be improved by the semiparametric ones. Other semiparametric approaches that use nonparametric and parametric estimations have been treated in regression with a linear component and a nonparametric component; compare Speckman (1988), Robinson (1988, 1989), and Cuzick (1992).

In Section 2 we study the problem of constructing prediction intervals adapted to Model (6). A new bootstrap approach, more general than that proposed by Thombs and Schucany (1990), is presented for autoregressive integrated (ARI) models. In Section 3 we present a simulation study comparing the already existing prediction techniques and the approach introduced in this article. This approach is proved to be competitive in such a simulation study as well as with real data related to ambient concentrations problems (as described in this introduction). Finally, we include a section of conclusions and an appendix with the proofs of the consistency of the nonparametric and parametric estimators given in (6) and of the bootstrap for ARI models used in our prediction intervals.

2. PREDICTION INTERVALS

2.1 Box–Jenkins Prediction Intervals

Because $\{e_l\}$ is not observable, it seems natural to work with the prediction $\dot{e}_t$ constructed from the estimated ARMA component recall, $\{\widehat{e}_{t-(n+k)+1} = Y_{t-(n+k)+1} - \widehat{\varphi}_n(Z_{t-(n+k)+1}), \ldots, \widehat{e}_{t-k} = Y_{t-k} - \widehat{\varphi}_n(Z_{t-k})\}$. The parameters of this ARMA model are consistently estimated under the same conditions as the ones in Theorem 1 in the Appendix. Moreover, taking into consideration the classical Box–Jenkins methodology, an asymptotic α-level prediction interval for Y_t can be constructed:

$$\widehat{\varphi}_n(Z_t) + \dot{e}_t \pm z_{\alpha/2}\left(\widehat{\sigma}^2 \sum_{j=0}^{k-1} \widehat{\pi}_j^2\right)^{1/2}, \quad (7)$$

where $z_{\alpha/2}$ represents the $1 - \alpha/2$ quantile of the standard normal, $\widehat{\sigma}^2$ is the classical estimation of the variance associated with the white-noise component of the ARMA series $\{e_l\}$, and the $\widehat{\pi}_j$ are the estimated coefficients of the polynomials π_j obtained from the relation

$$\pi(B) = \frac{\theta(B)}{\psi(B)(1 - B)^d},$$

where the θ and ψ coefficients are consistently estimated from the ARMA part $\{\widehat{e}_l\}$. The factor $(1 - B)^d$ is incorporated to cover the more general case in which $\{e_l\}$ follows an (autoregressive integrated moving average) ARIMA(q, d, s) model (see Wei 1990, p. 91, for further details).

2.2 Bootstrap Prediction Intervals

The prediction interval given in (7) is based on the critical assumption that $\{a_t\}$ is white noise; it behaves poorly when this assumption fails. An alternative interval is provided by the bootstrap methodology. For a simplified model in (5) without nonparametric component ($\varphi = 0$) and with $\{e_t\}$ following an AR(q) structure, Thombs and Schucany (1990) proposed a bootstrap mechanism to approximate the conditional distribution of e_t given $e_{t-k}, e_{t-k-1}, \ldots, e_{t-(n+k)+1}$.

Here we propose an adaptation of this mechanism that can be used more generally in ARI(q, d) models. Suppose that $\{e_t\}$ in (5) follows an ARI(q, d) structure with $\psi(B)(1 - B)^d e_t = a_t$, $\{a_t\}$ being noise. Clearly, the series $\tilde{e}_t = \nabla^d e_t$ follows an AR(q) model; $\tilde{e}^*_{t+i-k}(i = -n + 1 + d, \ldots, -q$ and $i = 1, \ldots, k)$ is obtained as done by Thombs and Schucany (1990), and the bootstrap series can be produced: $\{\tilde{e}^*_{t-(n-d)-k+1}, \ldots, \tilde{e}^*_{t-q-k}, \tilde{e}_{t-q-k+1}, \ldots, \tilde{e}_{t-k}, \tilde{e}^*_{t-k+1}, \ldots, \tilde{e}^*_t\}$. It is not difficult to show, by induction in d, that the $(k + d) \times (k + d)$ linear system

$$\nabla^d e^*_{t+i-k} = \tilde{e}^*_{t+i-k}, \qquad i = 1, \ldots, k,$$
$$e^*_{j-k} = e_{j-k}, \qquad j = t - d + 1, \ldots, t,$$

has a (unique) solution in e^*_j, $j = t - d + 1 - k, \ldots, t$. As a consequence, e^*_t can be expressed in terms of $\tilde{e}^*_t$, $\ldots, \tilde{e}^*_{t-k+1}, e_{t-k}, \ldots, e_{t-d-k+1}$. The bootstrap replication of this process many times produces the following approximate prediction interval (k lag) for e_t: $(z_t^{*(\alpha/2)}, z_t^{*(1-\alpha/2)})$, where $z_t^{*(\alpha/2)}$ and $z_t^{*(1-\alpha/2)}$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap distribution of e^*_t. Hence, under model (5), a semiparametric prediction interval for Y_t using the bootstrap prediction interval can be given by

$$\left(\hat{\varphi}_n(Z_t) + \hat{z}_t^{*(\alpha/2)}, \hat{\varphi}_n(Z_t) + \hat{z}_t^{*(1-\alpha/2)}\right), \tag{8}$$

where the bootstrap quantiles $\hat{z}^*$ are obtained from the ARMA component $\{\hat{e}_t\}$. The consistency of the interval $(z_t^{*(\alpha/2)}, z_t^{*(1-\alpha/2)})$ is a consequence of the consistency of the bootstrap for ARI models and is proved in Theorem 2 in the Appendix.

Although these bootstrap intervals when $\varphi \neq 0$ have not been shown to be consistent (this may be part of future research on this topic), we will see in Section 4 that its performance is competitive with Interval (7).

3. A SIMULATION STUDY

We conduct a comparative simulation study for the one-lag point prediction of three series $\{X_t\}$ using the Box–Jenkins methodology, the proposed semiparametric approach (5), and the pure nonparametric prediction. For every series, the study consists of repeating $N = 100$ times the following two steps:

1. A series $\{X_l, l = 1, \ldots, m = 200\}$ is generated as observations $301, 302, \ldots, 300 + m$ from the simulated model. Denote them as $X_1, X_2, \ldots, X_m$, with $X_0 = 0$.

2. Using $X_1, \ldots, X_m$, a new observation X_{m+1} is obtained $M = 1{,}000$ times.

The simulated models follow the structure in Model (5):

1. Series 1—ARMA(1, 1) model: $X_l = .7X_{l-1} + a_l - .4a_{l-1}$ with a_l white noise. This corresponds to the choice $Y_l = X_l, Z_l = X_{l-1}, e_l = a_l - .4a_{l-1}$, and $\varphi(Z_l) = .7Z_l$.

2. Series 2—pseudostationary model: $X_l = Z_l + e_l$ with $Z_l = 3 \sin \omega l$, $\omega = 2\pi/30$, and e_l with the same ARMA(1, 1) structure as the previous series.

3. Series 3—nonlinear AR model: $X_l = R(X_{l-1}) + \varepsilon_l$ with $R(x) = 19\pi/20 \sin x$, $x \in (0, \pi)$, and $\varepsilon_l \sim U[-\pi/20, \pi/20]$, iid. This corresponds to $Y_l = X_l$, $Z_l = X_{l-1}$, $e_l = \varepsilon_l$, and $\varphi = R$.

Observe that Series 1 is a pure Box–Jenkins model, but Series 2 is a slight perturbation of it and Series 3 is clearly nonlinear. For each of these models, we compare the one-lag predictor for the following three procedures:

1. The optimal estimated Box–Jenkins predictor [selected from among all possible ARIMA(q, d, s) models with $q \leq 6, d \leq 1, s \leq 6$] using International Mathematical and Statistical Libraries, Inc. (1991) routines.

2. The pure nonparametric predictor obtained estimating $E(X_{l+1} \mid X_l = x)$, using the kernel method. We considered the Gaussian kernel [with weight of type (3)]. The bandwidth $h_n (n = m - 1)$ is selected using a cross-validation method adapted to each point x; that is, $h_n = h_n(x)$ (Vieu 1991). The weight function for this cross-validation method is taken following the indications of Vieu (1991).

3. The semiparametric predictor obtained as described in Section 2. The nonparametric estimator is the predictor in 2, and the parametric estimator (for the residual series) uses the Box–Jenkins predictor in 1.

If we denote by $X^i_1, \ldots, X^i_{200}, i = 1, \ldots, N = 100$, each one of the three simulated models, and $\hat{X}^{ia}_{200+1}$, $\hat{X}^{ib}_{200+1}$, and $\hat{X}^{ic}_{200+1}$, each one of the three one-lag predictors, the methods are compared using the squared and absolute value criteria

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\left(X^{ij}_{200+1} - \hat{X}^{il}_{200+1}\right)^2 \tag{9}$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\left|X^{ij}_{200+1} - \hat{X}^{il}_{200+1}\right|, \tag{10}$$

where X^{ij}_{200+1} represents the observed value in the jth extension of the series $\{X_u\}_{u=1}^m$ (obtained in the corresponding repetition of Step 2), $j = 1, \ldots, M = 1{,}000$; $l = a, b$, or c. These criteria estimate the mean squared error (MSE) $E(\hat{X}_{l+1} - X_{l+1})^2$ and the mean absolute error (MAE) $E(|\hat{X}_{l+1} - X_{l+1}|)$, respectively.

Table 1. Mean Squared and Mean Absolute Errors

Bandwidth		Squa. SP			Squa. NP				Abs. SP			Abs. NP		
x factor	Squa. B–J	.5	1	1.5	.5	1	1.5	Abs. B–J	.5	1	1.5	.5	1	1.5
Model 1	1.0311	1.0456	1.0330	1.0305	1.0550	1.0555	1.0630	.8237	.8364	.8254	.8231	.8444	.8450	.8516
Model 2	1.2220	1.1695	1.1966	1.2561	1.2015	1.2239	1.3555	.9967	.9464	.9515	1.0295	.9752	.9946	1.1215
Model 3	.8002	.2881	.2869	.2905	.2891	.2596	.2857	.7919	.2732	.2731	.2761	.2742	.2453	.2712

The results obtained are shown in Table 1. The nonparametric and semiparametric estimations are carried out using three different bandwidths, obtained by multiplying the local cross-validation bandwidth by the factors .5, 1, and 1.5.

Figure 1 represents the variation of the MSE as a function of the bandwidth for the three methods.

The most relevant conclusions of this simulation study are:

1. The Box–Jenkins prediction is competitive when the structure is close to Box–Jenkins (Models 1 and 2). The nonparametric prediction behaves reasonably well for bandwidths with multiplicative factors close to 1, but gets worse as the factor increases (observed in simulation studies not included here). The semiparametric prediction is fairly good in any case, regardless of bandwidths and model, and is close to the best observed here for Models 1 and 2 (Box–Jenkins or nonparametric predictor).

2. The bandwidth selection is very important in the nonparametric method and in the semiparametric method. In general, the optimal local cross-validation bandwidth does not appear to be the optimal semiparametric one. The latter tends to be larger in situations like Model 1. This is natural because the nonparametric oversmoothing, present in the semiparametric methodology, is similar to the Box-Jenkins method.

3. In the nonlinear model, the Box-Jenkins method is definitely worse than the others. Figure 2 shows the three predictors for one of the simulated series of Model 3; similarly poor performance will be seen on real data in Section 4.

4. APPLICATION TO REAL DATA

We have applied the semiparametric prediction model in our design of a forecasting system implemented in the coal-fired power station in As Pontes, located in the northwest of Spain. Six tracking stations provided with automatic analyzers record SO_2 concentration levels and transmit them to the central laboratory of the station every 10 seconds. Every five minutes these data are averaged and the resulting value (together with the other 23 analogously obtained in the last two hours) is used to produce the *two-hour mean.* These two-hour means are the values to be controlled according to the Spanish laws. With the resources available in the power station at As Pontes, it takes about half an hour after the decision to intervene in the combustion process to achieve the required reduction. Because a new datum is recorded every five minutes, we have to predict the concentration level six times ahead and, according to that prediction, decide whether or not to intervene. Speed and accuracy of the forecast are very important.

Our experience with series of this type indicates that it can suddenly and sharply increase; these abrupt changes (episodes) are usually quite separated in time. All of the peculiarities mentioned previously should be considered when designing the forecasting model.

These characteristics suggest that Box–Jenkins methodology would not predict well. In Figure 3 we show the predictions (always six times ahead) along a real episode (August 11, 1990) using Box–Jenkins methodology, along with the forecastings (dotted line) and the observed series (thick solid line). The discontinuities of the dotted line are because sometimes it is not possible to find a suitable

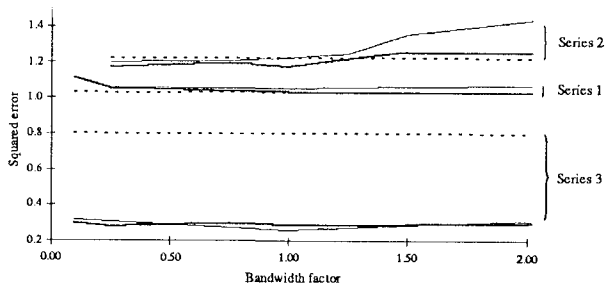

Figure 1. Unconditional Mean Squared Error—Series 1, ARMA(1, 1) Model; Series 2, Pseudostationary Model; Series 3, Nonlinear AR Model: ----, Box–Jenkins; ——; Semiparametric; ——, Nonparametric.

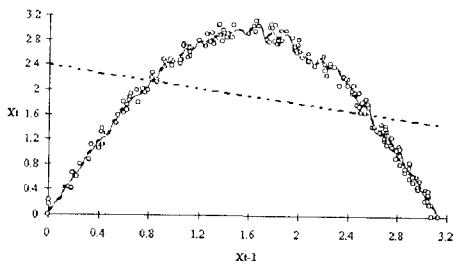

Figure 2. Estimation of the Autoregression Function $\varphi(X_{t-1}) = E(X_t \mid X_{t-1})$: ○, Data; ----, Box–Jenkins; — —, Semiparametric; – – –, Nonparametric.

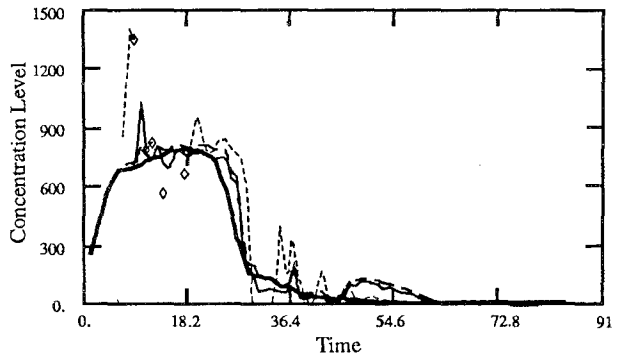

Figure 3. Prediction Using a Pure Box–Jenkins Model, the Nonparametric Methodology, and the Semiparametric Model: ———, Real; ———, Semiparametric; – – –, Nonparametric; - - - -, Box–Jenkins.

ARIMA model with a moderate number of parameters (we allow no more than two differences and no more than five parameters in the AR or the MA parts). The results are not very satisfactory.

In view of the drawbacks of the Box–Jenkins methodology, we estimated $E(X_{l+6} \mid X_l, X_{l-1})$ nonparametrically using the Nadaraya–Watson estimator [see (3)] with cross-validation bandwidth and Gaussian kernel. At every time, for modeling and inference, we consider the data recorded in the last six hours (72 observations). We call them the *active series* at that time. But now, the corresponding active series is not long enough to produce the prediction in every l. The reason is that, as we have already remarked, the episodes observed are too separated in time, so it is quite probable that, in a certain time, the active series does not contain an episode. Consequently, the nonparametric prediction can behave badly in the episodes when using only the active series in every time. Hence we decided to design a kind of memory for the prediction mechanism in such a way that it uses all of the experience accumulated up to that moment. This memory is embodied in what we call the *historical matrices*. They consist of 500 vectors of the form $(X_{\tau-1}, X_\tau, X_{\tau+6})$. Those vectors are grouped according to the value of their response component $X_{\tau+6}$, forming nine classes. At every time t, a new observation X_t is received. Then, the historical matrix is updated in the following way: The class to which the new vector (X_{t-7}, X_{t-6}, X_t) belongs is found; then, the oldest vector in such a class is replaced by (X_{t-7}, X_{t-6}, X_t). Now, if we use the historical matrix instead of the active series to compute the Nadaraya–Watson estimation of $E(X_{t+6} \mid X_t, X_{t-1})$, denoted by $\widehat{E}(X_{t+6} \mid X_t, X_{t-1}) = \widehat{\varphi}_n(X_t, X_{t-1})$, the nonparametric prediction improves remarkably (to avoid the "curse of dimensionality," we use only two explanatory variables). In Figure 3, we also show the nonparametric predictions (always six times ahead) using the sample of the historical matrix at every time t. The forecastings are displayed with a dashed line. The results are much better.

The nonparametric methodology seems to perform better than Box–Jenkins methods, but it is not completely satisfactory. In fact, if we perform a Ljung–Box test of model adequacy (Ljung and Box 1978) on the series $\widehat{W}_{t-64}, \ldots, \widehat{W}_t$, where $\widehat{W}_i := X_i - \widehat{E}(X_i \mid X_{i-6}, X_{i-7})$, for every i, the null hypothesis of a noise series is not always rejected. To address this issue, we first estimate $E(X_{t+6} \mid X_t, X_{t-1})$, then fit an ARIMA model to the series $\widehat{W}_{t-64}, \ldots, \widehat{W}_t$ (testing its adequacy with the Ljung–Box test) and obtain the Box–Jenkins prediction of $\widehat{W}_{t+6}$ (which would be 0 if the series of the $\widehat{W}_i$ is accepted to be noise by the Ljung–Box test). The final point prediction we propose is $\widehat{\varphi}_n(X_t, X_{t-1}) + \widehat{W}_{t+6}$ corresponding to Model (5), where $Z_t = (X_t, X_{t-1})$ and $e_t = W_{t+6}$.

In Figure 3 we also show the predictions (always six times ahead) using this semiparametric model (thin solid line). Comparing the three lines in Figure 3, we can see that the semiparametric model outperforms the pure Box–Jenkins approach, whereas it seems to be slightly better on average than the pure nonparametric approach. These statements are clearly confirmed in view of Table 2, which contains two measures of accuracy for the three different predictors,

$$(\mathrm{MSE})^{1/2} = \left((1/T) \sum_{l=t-T+1}^{t} \widehat{E}_l^2 \right)^{1/2}$$

and

$$\mathrm{MAE} = (1/T) \sum_{l=t-T+1}^{t} |\widehat{E}_l|,$$

where $\widehat{E}_l = X_{l+6} - \widehat{X}_{l+6}$ is the observed predictor error and $T = 79$ in our case. Both measures confirm that, along this episode, the semiparametric approach performs best, followed by the pure nonparametric approach and, far from the other two, by the pure Box–Jenkins approach. From these results it seems that the nonparametric estimation of $E(X_{l+6} \mid X_l, X_{l-1})$ captures the trend of the series very well (and, besides, makes it stationary), but the forecast can be improved: A further parametric modeling of the residuals may produce better predictions.

Table 3 displays the complexity (number of differences and total number of ARIMA parameters) of the ARIMA models that fitted the residual series or the observed series in the Box–Jenkins approach. We can see that, along this episode, in 9.41% of the cases the nonparametric prediction is sufficient (the estimated ARMA component turned

Table 2. Forecast Error

Model	*$MSE^{1/2}$*	*MAE*
Nonparametric	72.41	44.57
Semiparametric	67.64	40.99
Pure Box–Jenkins	171.55	88.38

Table 3. Number of Differences and Total Number of ARIMA Parameters

	Pure Box–Jenkins	*Semiparametric*
Differences		
0	.00%	93.50%
1	83.75%	6.50%
2	16.25%	.00%
Parameters		
White noise	.00%	9.41%
1 or 2	.00%	16.48%
3 or 4	14.12%	67.06%
More that 4	80.00%	7.05%
Out of range	5.88%	.00%

out to be noise). Moreover, when a further ARIMA model is needed, the number of parameters in such a model is four or less in more than 83% of the cases, whereas, when adopting a pure Box–Jenkins approach, only in 14.12% of the cases were four or less parameters enough. Regarding differencing, in the semiparametric approach, 93.5% of the time no differences were necessary in the parametric part of the model, whereas 100% of the time some differences were necessary when using the pure Box–Jenkins approach. Similar results were obtained in other episodes. In summary, it seems that the nonparametric part helps to capture the trend of the series and to find appropriate simpler ARIMA models for the corresponding residuals.

In Figure 4 we present the Box–Jenkins and bootstrap prediction intervals ($\alpha = .05$) obtained in every time as described in Section 2.

We can clearly see in Figure 4 that the bootstrap intervals perform better than the classical ones in this episode (similar results were observed in other episodes). This visual feeling is confirmed in Table 4, which shows how the bootstrap intervals outperform the classical ones (along this episode): The latter are too conservative, whereas the former are shorter and have a coverage rate much closer to the theoretical coverage (95%).

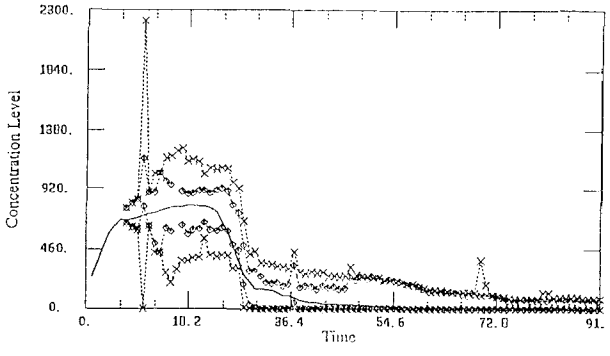

Figure 4. Classical and Bootstrap Confidence Intervals: ———, Real; ×----×, Box–Jenkins; ◇----◇, Bootstrap.

Table 4. Confidence Intervals (95%)

Intervals	*Length*	*Coverage*
Classical	343.43	100%
Bootstrap	196.94	94.87%

5. CONCLUSIONS

In this article, a new k-lag prediction system is presented. Such a system is based on a semiparametric mechanism that consists of the sum of a nonparametric prediction and the parametric Box–Jenkins prediction performed on the ARMA component of the series. Under quite general conditions (see Theorem 1 in the Appendix) the nonparametric and the parametric predictions are consistent. The behavior of this new system, when applied to a series $\{X_l\}$, is competitive in comparison with the Box–Jenkins and the nonparametric methodologies, as can be deduced from the simulation study in Section 3. Although the smoothing parameter is important in the semiparametric model, its importance is smaller than it is in the nonparametric estimation, as can be seen in Figure 1, where a higher sensitivity to the bandwidth is observed in the nonparametric method. One of the advantages of our procedure, observed in the simulation study as well as in the application to real data, is that it finds simpler predictive models than the pure Box–Jenkins procedure, which often has to use ARIMA models with many parameters (recall, for instance, Table 3) and a nontrivial MA part. This last point is very important because, to predict using bootstrap prediction intervals [of type (8)], one is forced to use an ARI model in the residual series.

It is important to remark that the environmental analysis performed in this article includes modern statistical techniques, such as curve estimation and bootstrap inference. In other recent environmetrical work, such as that of Uri (1991) and Hernández, Martín, and Valero (1992) about analysis of water quality and air quality, respectively, the statistical methodology consists of more classical tools such as Box–Jenkins models and Kalman filters.

ACKNOWLEDGMENTS

This article has improved remarkably due to the comments of the editor, an associate editor, and two referees. This work was partially supported by Grant PTR91-0794 from the Spanish Ministry of Education and Grant XUGA20701B92 from the Government of Galicia.

APPENDIX: MAIN PROOFS

A.1 Consistency of Nonparametric and Parametric Estimators

Under Model (5), the $(r + 1)$-dimensional series $\{(Z_l, Y_l)\}_{l\in\mathbb{Z}}$ is stationary. In this case, under (a) Lipschitz conditions on the function φ, (b) lower and upper boundedness

on the marginal density of Z_0 and the conditional densities of $Z_j \mid Z_0$ ($j \geq 1$) (assumed to exist), and (c) asymptotic independence conditions of α-mixing type with geometric rate [$\alpha(N) = O(\rho^N)$, where $N \to \infty$ and $0 < \rho < 1$], then consistency properties are proved for the nonparametric estimator of φ. More precisely, given a sample $\{(Z_i, Y_i)\}_{i=1}^n$, and using kernel weights

$$W_{ni}(z_0, (Z_1, Y_1), \ldots, (Z_n, Y_n)) = \frac{K\left(\frac{z_0 - Z_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{z_0 - Z_j}{h_n}\right)}$$

with uniform kernel K, the following optimal rates are derived for the nonparametric estimator $\widehat{\varphi}_n$:

$$\|\widehat{\varphi}_n - \varphi\|_\infty = O_{\mathbf{p}}\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2+r}}\right)$$

and

$$\|\widehat{\varphi}_n - \varphi\|_2 = O_{\mathbf{p}}\left(\left(\frac{1}{n}\right)^{\frac{1}{2+r}}\right)$$

(under boundedness of the ath absolute conditional moment of Y_0 given Z_0, for some $a > 2$); see Truong and Stone (1992) for details. The preceding norms $\| \ \|_\infty$ and $\| \ \|_2$ are computed on the support of Z_0. These rates require the appropriate choice of h_n.

Hence we have the consistency of the nonparametric estimator of φ in (5). We did not, however, treat the estimation of the parameters of the residual ARMA model yet. To do that, being more precise in Model (5), suppose that the series $\{e_l\}$ has an ARMA(q, s) structure: $\psi(B)e_l = \theta(B)a_l$, where ψ and θ are polynomials without common roots and all of the roots of ψ lie outside of the unit circle. We have the following result:

Theorem 1. Provided that the nonparametric estimator of φ is uniformly consistent [$\|\widehat{\varphi}_n - \varphi\|_\infty = o_{\mathbf{P}}(1)$], then $\|(\widehat{\theta}, \widehat{\psi}) - (\theta, \psi)\|_2 \to 0$ in probability, where $(\widehat{\theta}, \widehat{\psi})$ are the $(q + s)$-dimensional estimators of $(\theta_1, \ldots, \theta_s, \psi_1, \ldots, \psi_q) = (\theta, \psi)$ performed with the estimated ARMA component of the series $\widehat{e}_l = Y_l - \widehat{\varphi}_n(Z_l)$.

The estimation of (θ, ψ) is carried out using any mechanism that would be consistent if the estimation were made using the unobserved series $\{e_l\}$. For details about several consistent methods of estimating (θ, ψ) using $\{e_l\}$, see Brockwell and Davis (1991, chap. 8). Because they are based on the sample autocovariance function of the unobserved series

$$\overline{\gamma}_n(j) = \frac{1}{(n-j)} \sum_{l=t-n+1-k}^{t-k-j} e_l e_{l+j}, \qquad j = 0, 1, 2, \ldots,$$

to prove the consistency we should demonstrate that the estimated autocovariance function

$$\widehat{\gamma}_n(j) = \frac{1}{(n-j)} \sum_{l=t-n+1-k}^{t-k-j} \widehat{e}_l \widehat{e}_{l+j}, \qquad j = 0, 1, 2, \ldots,$$

satisfies that $\overline{\gamma}_n(j) - \widehat{\gamma}_n(j) = o_{\mathbf{P}}(1)$, $j = 0, 1, 2, \ldots$. This is proved later.

Because

$$\begin{aligned}\widehat{e}_l \widehat{e}_{l+j} = e_l e_{l+j} &- e_{l+j}(\widehat{\varphi}_n(Z_l) - \varphi(Z_l)) \\ &- e_l(\widehat{\varphi}_n(Z_{l+j}) - \varphi(Z_{l+j})) \\ &+ (\widehat{\varphi}_n(Z_l) - \varphi(Z_l)) \times (\widehat{\varphi}_n(Z_{l+j}) - \varphi(Z_{l+j})), \\ &l = t - n + 1 - k, \ldots, t - k - j,\end{aligned}$$

and taking into account the uniformly consistency of $\widehat{\varphi}_n$ ($\|\widehat{\varphi}_n - \varphi\|_\infty = o_{\mathbf{P}}(1)$) and that $E(\frac{1}{n}\sum_{l=t-n+1-k}^{t-k} |e_l|) = E(|e|) < \infty$; $\frac{1}{n}\sum_{l=t-n+1-k}^{t-k} |e_l| = o_{\mathbf{P}}(1)$, $\overline{\gamma}_n(j) - \widehat{\gamma}_n(j) = o_{\mathbf{P}}(1)$, $j = 0, 1, 2, \ldots$, as stated.

There are some possible estimators for (θ, ψ); we consider the estimators based on the Durbin–Levinson algorithm. From the assumptions in this theorem, the unobservable ARMA(q, s) series $\{e_l\}$ satisfies

$$e_l = \sum_{j=0}^{\infty} \eta_j a_{l-j}$$

with

$$\eta_j - \sum_{0<k\leq j} \psi_k \eta_{j-k} = \theta_j, \qquad 0 \leq j < \max(q, s+1)$$

and

$$\eta_j - \sum_{0<k\leq q} \psi_k \eta_{j-k} = 0, \qquad j \geq \max(q, s+1).$$

Using the innovation algorithm, $(\widehat{\theta}_{m1}, \ldots, \widehat{\theta}_{mm})$ with $m = o(n^{1/3})$ gives the best linear predictor $\theta_{m1}e_m + \cdots + \theta_{mm}e_1$ of e_{m+1} (see Brockwell and Davis 1991, p. 246). Such estimators depend on the autocovariance function $\overline{\gamma}_n(j)$ [Brockwell and Davis 1991, formulas (5.2.3) or (8.3.2)]. Substituting the estimators $\widehat{\theta}_{m1}, \ldots, \widehat{\theta}_{m(q+s)}$ instead of $\eta_1, \ldots, \eta_{q+s}$ in the preceding equations, we have consistent estimators for $(\theta_1, \ldots, \theta_s, \psi_1, \ldots, \psi_q)$ (cf. Brockwell and Davis 1991, p. 253).

Assuming some differentiability conditions for φ and using a (nonparametric) local polynomial regression estimator $\widehat{\varphi}_n$ of a large enough degree g ($g = 0$ leads to a uniform kernel estimator), stronger properties, such as the $\sqrt{n}$-consistency, $\|(\widehat{\theta}, \widehat{\psi}) - (\theta, \psi)\|_2 = o_{\mathbf{P}}(n^{-1/2})$, can be obtained (see Truong 1992).

A.2 Consistency of the ARI Bootstrap

As an extension of Thombs and Schucany's (1990) result, the following theorem is stated for ARI models:

Theorem 2. Consider the model $\psi(B)(1-B)^d e_l = a_l$ with $E(a_l) = 0$ and $E(|a_l|^\alpha) < \infty$ for some $\alpha > 2$. Then $e_t^* \to e_t$ in distribution when $n \to \infty$ almost surely; that is,

$$\begin{aligned}\lim_{n\to\infty} \big(&P_*(e_t^* \leq z \mid e_{t-k}, \ldots, e_{t-n-k+1}) \\ &- P(e_t \leq z \mid e_{t-k}, \ldots, e_{t-n-k+1})\big) = 0\end{aligned}$$

almost surely.

Proof. An immediate consequence of part (b) of theorem 3.1 of Thombs and Schucany for the series $Y_l = (1 - B)^d e_l$ is that, for almost every sample path $\{Y_{t-k}, \ldots, Y_{t-n+1+d-k}\}$, it is verified that $Y_t^* \to Y_t$ (in distribution) when $n \to \infty$.

We will prove further that, almost surely (in the same sense as before)

$$\begin{pmatrix} Y^*_{t+1-k} \\ Y^*_{t+2-k} \\ \cdots \\ Y^*_t \end{pmatrix} \to \begin{pmatrix} Y_{t+1-k} \\ Y_{t+2-k} \\ \cdots \\ Y_t \end{pmatrix} \tag{A.1}$$

in distribution when $n \to \infty$.

Observe that e_t^* (respectively e_t) can be represented as a finite linear combination of $Y^*_{t+1-k}, \ldots, Y_t^*$ ($Y_{t+1-k}, \ldots, Y_t$, respectively) and $e_{t-d+1-k}, \ldots, e_{t-k}$ (for example, for $k = 2$ and $d = 3$, $e_t^* = Y_t^* + 3Y^*_{t-1} + 6e_{t-2} - 8e_{t-3} + 3e_{t-4}$), where the coefficients of such a combination are the same both for e_t^* and e_t. Note also that $\{e_{t-n+1-k}, \ldots, e_{t-k}\}$ and $\{e_{t-n+1-k}, \ldots, e_{t-n+d-k}, Y_{t-n+d+1-k}, \ldots, Y_{t-k}\}$ generate the same σ algebra. Hence, taking into account the representations of e_t^* and e_t and the fact that both random vectors in (A.1) conditionally on $e_{t-n+1-k}, \ldots, e_{t-n+d-k}, Y_{t-n+d+1-k}, \ldots, Y_{t-k}$ have the same distribution as they have conditionally on $Y_{t-n+d+1-k}, \ldots, Y_{t-k}$, we can assert that $e_t^* \to e_t$ (in distribution) when $n \to \infty$ along almost every path $\{e_{t-n+1-k}, \ldots, e_{t-k}\}$.

In the remainder, we prove Expression (A.1). To simplify the notation we consider the case $k = 2$. Because the series $\{Y_l\}$ follows an AR(q) model ($Y_l = \psi_1 Y_{l-1}) + \cdots + \psi_q Y_{l-q} + a_l$), it is easy to see that

$$\begin{aligned}(Y_{t-1}, Y_t) &= (\psi_1 Y_{t-2} + \cdots + \psi_q Y_{t-q-1}, (\psi_1^2 + \psi_2) \\ &\quad \times Y_{t-2} + (\psi_1\psi_2 + \psi_3) Y_{t-3} + \cdots \\ &\quad + (\psi_1\psi_{q-1} + \psi_q) Y_{t-q} + \psi_1\psi_q Y_{t-q-1} \\ &\quad + \psi_1 a_{t-1} + a_t) \\ &= (\psi_1 Y_{t-2} + \cdots + \psi_q Y_{t-q-1}, (\psi_1^2 + \psi_2) Y_{t-2} \\ &\quad + \cdots + \psi_1\psi_q Y_{t-q-1}) + ((1, \psi_1) a_{t-1} \\ &\quad + (0, 1) a_t) \\ &= \mathrm{I} + \mathrm{II}.\end{aligned}$$

In view of the resampling procedure proposed by Thombs and Schucany (1990),

$$\begin{aligned}(Y^*_{t-1}, Y^*_t) &= (\psi_1^* Y_{t-2} + \cdots + \psi_q^* Y_{t-q-1}, (\psi_1^{*2} + \psi_2^*) \\ &\quad \times Y_{t-2} + (\psi_1^*\psi_2^* + \psi_3^*) Y_{t-3} + \cdots \\ &\quad + (\psi_1^*\psi_{q-1}^* + \psi_q^*) Y_{t-q} + \psi_1^*\psi_q^* Y_{t-q-1} \\ &\quad + \psi_1^* a^*_{t-1} + a_t^*) \\ &= \mathrm{I}^* + \mathrm{II}^* + \mathrm{III}^*\end{aligned}$$

with $\mathrm{I}^* = (\psi_1^* Y_{t-2} + \cdots + \psi_q^* Y_{t-q-1}, (\psi_1^{*2} + \psi_2^*) Y_{t-2} + \cdots + \psi_1^*\psi_q^* Y_{t-q-1})$, $\mathrm{II}^* = ((1, \psi_1^*) - (1, \psi_1)) a^*_{t-1}$, and $\mathrm{III}^* = ((1, \psi_1) a^*_{t-1} + (0, 1) a_t^*)$.

A consequence of theorem 3.1(a) of Thombs and Schucany (1990) is that, almost surely, $(\psi_1^*, \ldots, \psi_q^*) \to (\psi_1, \ldots, \psi_q)$ (in $\mathbf{P}^*$), and then $I^* \to I$ (in $\mathbf{P}^*$) almost surely. Besides, because a^*_{t-1} and a_t^* are independent when conditioned to the sample $\{Y_{t-n+d-1}, \ldots, Y_{t-2}\}$ and $a^*_{t-1} \to a_{t-1}$ and $a_t^* \to a_t$ (both in distribution) almost surely (see Freedman 1985), then $\mathrm{II}^* \to 0$ (in $\mathbf{P}^*$) and $\mathrm{III}^* \to \mathrm{II}$ (in distribution) almost surely. Hence the proof is concluded.

[*Received February 1992. Revised February 1995.*]

REFERENCES

Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods*, New York: Springer-Verlag.

Cuzick, J. (1992), "Efficient Estimates in Semiparametric Additive Regression Models With Unknown Error Distribution," *The Annals of Statistics*, 2, 1129–1136.

Freedman, D. A. (1985), "On Bootstrapping Two-Stage Least Squares Estimates in Stationary Linear Models," *The Annals of Statistics*, 12, 827–842.

Hernández, E., Martín, F., and Valero, F. (1992), "Statistical Forecast Models for Daily Air Particulate Iron and Lead Concentrations for Madrid, Spain," *Atmospheric Environment*, 1, 107–116.

International Mathematical and Statistical Libraries, Inc. (1991), *STAT/LIBRARY* (version 2.0), Houston: Author.

Ljung, G. M., and Box, G. E. P. (1978), "On the Measure of Lack of Fit in Time Series Models," *Biometrika*, 65, 297–303.

Robinson, P. M. (1988), "Root n-consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

——— (1989), "Hypothesis Testing in Semiparametric and Nonparametric Models for Economic Time Series," *Review of Economic Studies*, 56, 511–534.

Speckman, P. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 3, 413–436.

Thombs, L. A., and Schucany, W. R. (1990), "Bootstrap Prediction Intervals for Autoregression," *Journal of the American Statistical Association*, 85, 486–492.

Truong, Y. K. (1992), "A Nonparametric Framework for Time Series Analysis," in *New Directions in Time Series Analysis*, eds. D. R. Brillinger et al., New York: Springer-Verlag, pp. 371–386.

Truong, Y. K., and Stone, C. J. (1992), "Nonparametric Function Estimation Involving Time Series," *The Annals of Statistics*, 1, 77–97.

Uri, N. D. (1991), "Water Quality, Trend Detection, and Casualty," *Water, Air, and Soil Pollution*, 59, 271–279.

Vieu, P. (1991), "Nonparametric Regression: Local Optimal Bandwidth Choice," *Journal of the Royal Statistical Society*, Ser. B, 2, 453–464.

Wei, W. (1990), *Time Series Analysis*, New York: Addison-Wesley.

Yakowitz, S. J. (1985), "Nonparametric Density Estimation, Prediction and Regression for Markov Sequences," *Journal of the American Statistical Association*, 339, 215–221.

——— (1987), "Nearest-Neighbour Methods for Time Series Analysis," *Journal of Time Series*, 2, 235–247.