Application of a Local Linear Autoregressive Model to BOD Time Series

Zongwu Cai and Ram C. Tiwari
Department of Mathematics
University of North Carolina
Charlotte, NC 28223

Abstract

In this paper, we analyze the biochemical oxygen demand data collected over two years from McDowell Creek, Charlotte, North Carolina, USA, by fitting an autoregressive model with time-dependent coefficients. The local linear smoothing technique is developed and implemented to estimate the coefficient functions of the autoregressive model. A nonparametric version of the Akaike information criterion is developed to determine the order of the model and to select the optimal bandwidth. We also propose a hypothesis testing technique, based on the residual sum of squares and F-test, to detect whether certain coefficients in the model are really varying or whether any variables are significant. The approximate null distributions of the test are provided. The proposed model has some advantages such as it is determined completely by data, it is easily implemented and it provides a better prediction.

Key Words: Akaike information criterion; bandwidth selection; goodness-of-fit; local linear fitting; order determination; varying-coefficient models.

1 Introduction

Each year millions of dollars are spent by various agencies to monitor the quality of water as the water pollution continues to be one of many environmental problems that mankind is facing today. The thousands of tons of organic pollutants and wastes discharged by industries and municipalities in streams, lakes and rivers greatly affect the quality of water and the various aquatic processes. Recognizing the effects of water pollution, community and government agencies have set up guidelines for the amount and type of organic pollutants that can be discharged daily in these water resources. It is generally agreed by scientists and regulators that the biochemical oxygen demand (BOD) and dissolved oxygen (DO), both measured in parts per million (ppm), are important indicators in assessing the level of water pollution.

Observed time series from BOD or DO data are often modeled in terms of regression models or state space models corresponding to trend, seasonal effects and an irregular noise components. See, for example, Papadopoulos *et al.* (1991), Papadopoulos and Tiwari (1994), Tiwari and Dienes (1994), and Tiwari, Yang and Zalkikar (1996, hereafter, referred to as TYZ). Sometimes, the observed series have almost identical seasonal components but it is more often the case that the seasonal factors vary from time to time, say year to year, in some relatively smooth fashion. In addition, the trends are often smooth but can not be fitted well by straight lines or higher order polynomials because the character of the smooth variation changes over time. To illustrate, Figure 1 shows the time series plot of the BOD data collected from February, 1982 to February, 1984 from

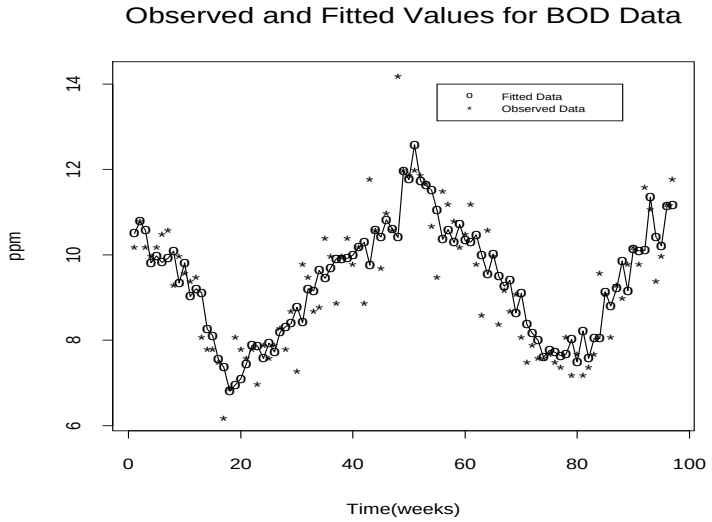

Figure 1: *The time series plots of the fitted values (indicated by "o") and the true observations (denoted by "∗") for BOD Data.*

McDowell Creek, Charlotte, North Carolina, USA. The data consist of 109 weekly measurements of BOD levels (measured in ppm).

Based on several assumptions, various methods for model fitting and predicting the levels of BOD and DO have been proposed and studied. Padgett and Rao (1979) utilized a nonparametric approach for estimating the joint probability density functions of BOD and DO along a stretch

of stream, Fuller and Tsokos (1971) performed a special time series analysis of DO data, and Papadopoulos *et al.* (1991) analyzed the BOD data $\{Y_t\}$ by fitting the sine wave (regression) model

$$Y_t = \mu + \alpha \sin(\beta\, t - \gamma) + \varepsilon_t, \tag{1.1}$$

where μ, α, β, γ and $\{\varepsilon_t\}$ are, respectively, the overall mean, amplitude, period, phase shift and model errors, and they employed the empirical bootstrap and the Bayesian bootstrap techniques to obtain the parameter estimators. Furthermore, a Bayesian approach to analyze BOD data was first used by Papadopoulos and Tiwari (1994). Tiwari and Dienes (1994) extended model (1.1) by allowing the parameters μ and α to be time dependent random variables by using a Kalman filter approach. More precisely, they considered the following state space model

$$Y_t = \mu_t + \alpha_t \cos(2\pi\, \omega\, t + \phi) + \varepsilon_t \tag{1.2}$$

with $\mu_t = \mu_{t-1} + \eta_t$ and $\alpha_t = \alpha_{t-1} + v_t$, where the errors ε_t, η_t and v_t are independent and normally distributed with zero mean and known variances. Recently, TYZ (1996) relaxed the assumption on the error variances and used a Gibbs sampler technique to fit model (1.2). For the aforementioned data set, TYZ (1996) used the first 100 data points to estimate the parameters via a Kalman filter and by using the Gibbs sampler and the last 9 observations were left to check the predictive performance of the model, so do we in Section 3. They showed that model (1.2) gives a better fit than the previously used models. However, the algorithm of using the Gibbs sampler requires an extensive computation and it does not provide accurate predictions of the future observations. See TYZ (1996, p.576) for more on difficulties of using the Gibbs sampler to do prediction.

We now propose a data-analytic approach to analyze the BOD data by using the following autoregressive (AR) model with time-dependent coefficients

$$Y_t = g_0(t) + \sum_{j=1}^{p} g_j(t)\, Y_{t-j} + \varepsilon_t, \tag{1.3}$$

where $\{g_j(\cdot)\}$ are unknown functions and the random errors $\{\varepsilon_t\}$ satisfy $E(\varepsilon_t \,|\, Y_{t-1}, \ldots, Y_{t-p}) = 0$ and $\mathrm{Var}(\varepsilon_t \,|\, Y_{t-1}, \ldots, Y_{t-p}) = \sigma^2 > 0$, unknown. $g_0(\cdot)$ is regarded as a time trend component and $Y_{t-j} (1 \le j \le p)$ are lagged variables which change over time in accordance with the modulating functions $\{g_j(\cdot)\}$, which describe how the lagged variable changes over time. Model (1.3) is clearly an extension of model (1.1).

In Section 2, we first develop a local linear smoothing technique to estimate the coefficient functions $\{g_j(\cdot)\}$ in model (1.3). We also propose a nonparametric version of the Akaike information criterion (AIC) to determine the order p of the model. Finally, we derive a hypothesis testing procedure, based on the residual sum of squares and F-test, to detect whether certain coefficient functions $\{g_j(\cdot)\}$ are really varying or whether any variables are statistically significant in the model. The approximate null distributions of the test are also provided.

Section 3 reports the detailed analysis of the aforementioned BOD data using the methodology described in Section 2. We use model (1.3) with $p = 3$, which is determined by the data using AIC, to fit the dataset. We show that model (1.3) outperforms model (1.2) in terms of one-step prediction.

More importantly, our method is particularly easy to understand and implement. Indeed, model (1.3) provides an approach to the identification of non-linear time series which require less structure to be imposed on the data and is arguably more helpful in exploring a set of data to understand the nature of the response surface. Furthermore, within the framework of (1.3), the detailed form of model is determined by data, which reduces the bias of fitting automatically and improves the post-sample predictions.

2 Modeling Methods

The AR model (1.3) is a special case of a more general model, known as the varying-coefficient model introduced by Hastie and Tibshirani (1993), which assumes the form

$$Y_t = \sum_{j=0}^{p} g_j(\mathbf{U}_t)\, X_{tj} + \varepsilon_t, \tag{2.1}$$

where $\mathbf{U}_t$ is a covariate vector in the k-dimensional Euclidean space, $\Re^k (k \geq 1)$, $\{g_j(\cdot)\}$ are unknown functions, and $\{\varepsilon_t\}$ are the random errors satisfying $E(\varepsilon_t \,|\, \mathbf{U}_t,\, \mathbf{X}_t) = 0$ and $\mathrm{Var}(\varepsilon_t \,|\, \mathbf{U}_t,\, \mathbf{X}_t) = \sigma^2$. Here, $\mathbf{X}_t = (X_{t0},\, \ldots,\, X_{tp})^T$ is also a covariate vector and T denotes the transpose of a matrix or vector. In a pure time series context, both $\mathbf{U}_t$ and $\mathbf{X}_t$ consist of some lagged values of $\{Y_t\}$. The varying-coefficient models are a simple and useful extension of classical linear models for iid errors and AR models in time series contexts. The varying-coefficient model allows the coefficients to change with a covariate or time, thus it increases the flexibility of the structure of the fitted models without suffering from the "curse of dimensionality". Model (2.1) includes some well-known nonlinear time series models such as the functional-coefficient autoregressive models and the threshold autoregressive models, studied by Tong (1990), Chen and Tsay (1993) and Cai, Fan and Yao (1998). See these references for more details on these models.

The focus, in the remainder of this section, is on model (2.1). For simplicity, we consider only the case where $\mathbf{U}$ in (2.1) is one-dimensional. Extension to multivariate $\mathbf{U}$ involves fundamentally no new ideas.

2.1 Local linear estimation

We use local linear regression method due to its nice properties such as high statistical efficiency (in an asymptotic minimax sense), design adaptation, automatic good boundary behavior (Fan and Gijbels, 1996). Suppose that the second derivative of $g_j(\cdot)$ exists and is continuous. For each given point u_0, we approximate function $g_j(\cdot)$ locally by linear function $g_j(u) \approx a_j + b_j\,(u - u_0)$ for u in a neighborhood of u_0. Note that a_j and b_j depend on u_0. Based on a random sample $\{U_t,\, X_{t0},\, \ldots,\, X_{tp},\, Y_t\}_{t=1}^n$, we consider the weighted least-squares problem to minimize

$$\sum_{t=1}^{n} \left[Y_t - \sum_{j=0}^{p} \{a_j + b_j\,(U_t - u_0)\}\; X_{tj} \right]^2 K_h(U_t - u_0), \tag{2.2}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a given kernel function, and $h = h_n > 0$ is the bandwidth. The kernel function $K(\cdot)$ plays the role to down-weight the contributions of a data point away from u_0.

Minimizing (2.2) with respect to $\{a_j\}$ and $\{b_j\}$ gives the estimator of $\{g_j(u_0)\}$. Let $\widehat{g}_j(u_0)$ be the estimator of $g_j(u_0)$. Then, it can be shown that

$$\widehat{g}_j(u_0) = e_{j,2(p+1)}^T \left(\widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^T \mathbf{W} \mathbf{Y}, \qquad j = 0, \ldots, p,$$

where $e_{j,m}$ is the $m \times 1$ unit vector with 1 at the $(j+1)$-th position,

$$\widetilde{\mathbf{X}} = \widetilde{\mathbf{X}}(u_0) = \begin{pmatrix} X_{10} & \cdots & X_{1p} & X_{10}(U_1 - u_0) & \cdots & X_{1p}(U_1 - u_0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_{n0} & \cdots & X_{np} & X_{n0}(U_n - u_0) & \cdots & X_{np}(U_n - u_0) \end{pmatrix},$$

$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, and $\mathbf{W} = \mathbf{W}(u_0) = \text{diag}\{K_h(U_1 - u_0), \ldots, K_h(U_n - u_0)\}$, and that the fitted value vector $\widehat{\mathbf{Y}} = \left(\widehat{Y}_1, \ldots, \widehat{Y}_n\right)^T$ with $\widehat{Y}_t = \sum_{j=0}^p \widehat{g}_j(U_t) X_{tj}$ can be expressed as

$$\widehat{\mathbf{Y}} = \mathbf{S}^* \mathbf{Y},$$

where

$$\mathbf{S}^* = (\mathbf{S}_1, \ldots, \mathbf{S}_n)^T, \qquad \mathbf{S}_t = \mathbf{A}(U_t)^T \mathbf{X}_t, \tag{2.3}$$

and

$$\mathbf{A}(u_0) = (\mathbf{I}_p, \mathbf{0}) \left(\widetilde{\mathbf{X}}(u_0)^T \mathbf{W}(u_0) \widetilde{\mathbf{X}}(u_0)\right)^{-1} \widetilde{\mathbf{X}}(u_0)^T \mathbf{W}(u_0).$$

Note that $\mathbf{S}^*$ is called a "linear smoothing operator" in literature (see Hastie and Tibshirani, 1990).

2.2 Order determination and bandwidth selection

After fitting the model, of importance is to determine the order-parameter p in model (2.1). There are numerous papers on stepwise selection of variables in the parametric regression and time series literature. Most of these are concerned with linear models. To our knowledge, a nonparametric approach to this problem has not been attempted. To select the order-parameter p in model (2.1), regard (2.2), for the fixed grid point u_0, as a least-squares problem for a linear regression model, and apply the bias-corrected AIC, due to Hurvich and Tsai (1989). The procedure is described below.

In the parametric regression and autoregressive time series contexts, Hurvich and Tsai (1989) derived the bias-corrected AIC, called AIC_C in literature, which assumes the form

$$\text{AIC}_C = \log \widehat{\sigma}^2 + \frac{n + n_p}{n - (n_p + 2)}, \tag{2.4}$$

where n_p denotes the number of parameters and $\widehat{\sigma}^2$ is the mean squared errors (MSE) under the fitted model. It has been suggested that we try the AIC_C, as recommended by Brockwell and Davis (1991), section 9.3, which penalizes extra parameters for larger values of n_p. The natural question is how to define n_p in the nonparametric setting. Here, we define $\text{tr}(\mathbf{S}^*)$ to be the number of parameters involved in the model fitting, called the degrees of freedom (*df*) in literature for the linear smoothing operator $\mathbf{S}^*$ given in (2.3). Indeed, this definition of *df* is used in Hastie and Tibshirani (1990, 1993) and is also popular in the smoothing spline literature. There are two other popular definitions of *df* for a linear smoothing operator besides $\text{tr}(\mathbf{S}^*)$, namely, $n - \text{tr}(2\,\mathbf{S}^* - \mathbf{S}^*\,\mathbf{S}^{*T})$

and $\mathrm{tr}(\mathbf{S}^*\mathbf{S}^{*T})$. These definitions, as well as $\mathrm{tr}(\mathbf{S}^*)$, can be motivated by analogy with the linear regression model and are useful for different purposes. For the nonparametric setting, $n_p = \mathrm{tr}(\mathbf{S}^*)$ would be very large since the parameter space is functional space with infinite dimension. Therefore, the AIC_C is particularly suitable for this case. We now propose the following nonparametric version of the AIC_C for order selection for model (2.1)

$$\mathrm{AIC}_C(h,\, p) = \log\{\mathrm{RSS}\} + \frac{n + \mathrm{tr}(\mathbf{S}^*)}{n - [\mathrm{tr}(\mathbf{S}^*) + 2]}, \tag{2.5}$$

where $\mathrm{RSS} = \sum_{t=1}^{n} \left(Y_t - \widehat{Y}_t\right)^2 = \mathbf{Y}^T (\mathbf{I} - \mathbf{S}^*)^T (\mathbf{I} - \mathbf{S}^*)\, \mathbf{Y}$. For notational convenience, henceforth, the subscript is dropped from $\mathrm{AIC}_C(h,\, p)$.

The resulting model (2.1) is chosen in such a way that $\mathrm{AIC}(h,\, p)$ is minimized with respect to p for each given bandwidth from a wide range. Based on our limited experience, the choice of bandwidth is not sensitive in the step of selecting the order-parameter p. In other words, for each given bandwidth from a wide range, $\mathrm{AIC}(h,\, p)$ should be convex in p. Note that the formula (2.5) can be used to select bandwidth. See Hurvich, Simonoff and Tsai (1998) for detail discussions for the nonparametric regression problems under iid setting. In our implementation, we have also made use of the formula (2.5) to select the bandwidth for the given model. Also, note that (2.5) has a connection to generalized cross-validation, which is also discussed in Hurvich, Simonoff and Tsai (1998).

We now outline the algorithm to select the order and bandwidth as follows:

Step 1: For each given bandwidth from a wide range with certain increment, say, 0.05, compute $\mathrm{AIC}(h,\, p)$ as a function of order-parameter p. Minimizing $\mathrm{AIC}(h,\, p)$ with respect to p, gives the best selection of order p, say, $\widehat{p}$.

Step 2: For given $p = \widehat{p}$ determined in Step 1, use (2.5) again to select the optimal bandwidth $\widehat{h}_{opt}$.

Note that (2.5) can be also used as a criterion to search for the lagged-parameter d if $U_t = Y_{t-d}$ and $X_{tj} = Y_{t-j}$, which was discussed in detail by Cai and Tiwari (1998).

2.3 Testing hypothesis

Having fitted the model, we would like to test whether the coefficient functions are really varying or whether any particular variables are significant in the model. The test of constant coefficient functions is of particular interest because it indicates whether the linear AR model is adequate. Furthermore, it is often of interest to test whether any variable can be deleted from the model. For simplicity of discussion, we only focus on the hypothesis testing problem

$$H_0 :\ g_0(u) \equiv \beta_0, \cdots, g_p(u) \equiv \beta_p,$$

though the technique continues to apply to other testing problems. A useful approach is based on the residual sum of squared (RSS), which is F-test, defined by

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(\nu_1 - \nu_0)}{\mathrm{RSS}_1/(n - \nu_1)}, \tag{2.6}$$

where RSS_0 and RSS_1 are the RSS, computed from the null and alternative hypotheses with the degrees of freedom ν given by $\nu = \mathrm{tr}(\mathbf{S}^* - \mathbf{S}^*\mathbf{S}^{*T})$ with $\mathbf{S}^*$ being a linear smoothing operator defined in (2.3). The approximate distribution of the test statistic F is $F(\nu_1 - \nu_0,\, n - \nu_1)$ (see Hastie and Tibshirani, 1990, p.66). The more refined approximation can be found in Cleveland and Devlin (1988) by applying a two-moment correction to both the numerator and denominator.

Note that since $\mathrm{tr}(\mathbf{S}^*)$ is easy to compute, whereas $\mathrm{tr}(\mathbf{S}^*\mathbf{S}^{*T})$ is not. The approximation (see Hastie and Tibshirani, 1990, p.305) $\mathrm{tr}(2\mathbf{S}^* - \mathbf{S}^*\mathbf{S}^{*T}) \approx 1.25\,\mathrm{tr}(\mathbf{S}^*) - 0.5$ is used here. We have made use of this approximation in the present work as well.

In the parametric context, it is well-known that under some regularity conditions, F defined in (2.6) is F-distribution with degrees of freedom $(\nu_1 - \nu_0)$ and $(n - \nu_1)$. For the nonparametric setting, the effective number of parameters would be very large since the parameter space is functional space with infinite dimension. Recalling the procedure of local least-squares method, the effective number of parameters used here to estimate the coefficient functions is much less than $(p+1) \times n_0$ (denoting the number of grid points), but it may be still a large number. Therefore, the standardization f of F given by

$$f = (F - 1)\sqrt{\frac{\nu_1 - \nu_0}{2}} \tag{2.7}$$

should be used, instead of F itself. The distribution of f can be approximated by the standard normal distribution $N(0,\, 1)$.

3 Modeling The BOD Data

Now we use our proposed procedures to analyze the BOD data mentioned in Section 1. As mentioned in the introduction, the data consist of 109 weekly measurements of BOD levels, and the time series plot of the data set is illustrated in Figure 1. We compare the predictive performance of the proposed model (1.3) with TYZ's model (1.2) based on the one-step ahead predictor, which is defined as follows for Y_{t+1} in model (1.3)

$$\widehat{Y}_{t+1} = \widehat{g}_0(t+1) + \widehat{g}_1(t+1)\, Y_t + \cdots + \widehat{g}_p(t+1)\, Y_{t-p+1}, \tag{3.1}$$

where $\widehat{g}_j(\cdot)$ is computed based on the first 100 observations. We predict the BOD level for the last nine observations $Y_{101}, \ldots, Y_{109}$. For estimation and prediction, the Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ was used.

We used the procedures described in Section 2.2 to select the order of p and the optimal bandwidth. First, we chose 9 initial bandwidths ranging from 0.20 to 0.60 with increment 0.05 and fitted the model with order p from 1 to 5 for each bandwidth. Figure 2 depicts the AIC as a function of p for bandwidth $0.20 \le h \le 0.60$ and it gives the evidence that the AIC(p) is minimized at $\widehat{p} = 3$. Therefore, the best model for the BOD data assumes the form as follows

$$Y_i = g_0(t_i) + g_1(t_i)\, Y_{i-1} + g_2(t_i)\, Y_{i-2} + g_3(t_i)\, Y_{i-3} + \varepsilon_i, \tag{3.2}$$

where $t_i = i/T,\ \ i = 1, \ldots, T$, and $T = 109$.

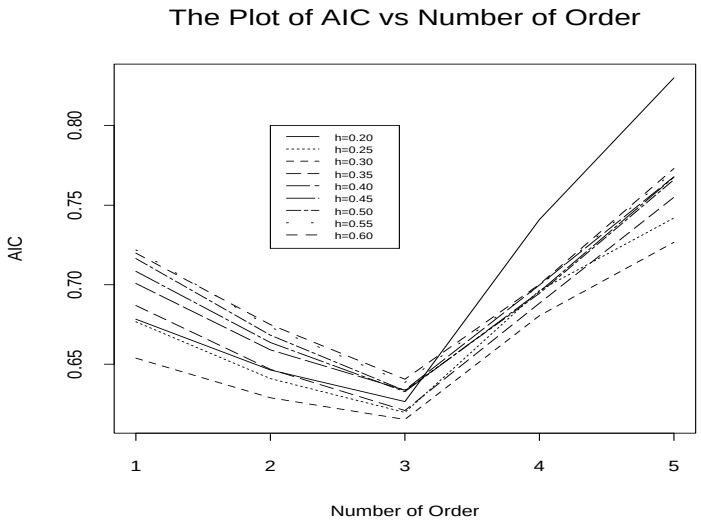

Figure 2: *The plot of AIC against the number of order for BOD Data.*

Next, for model (3.2), we applied the AIC to select the optimal bandwidth and the resulting bandwidth $\widehat{h}_{opt} = 0.3$ (see Figure 2). The average sum of squared errors $\widehat{\sigma}^2$ is 0.4573, the *df* used to fit the model is 14.9141 and the AIC is 0.6150. The estimated coefficient functions are summarized in Figure 3, which shows that the coefficient functions $g_j(\cdot)$ $(0 \leq j \leq 3)$ vary with time. The fitted values from model (3.2) and the true values of the first 100 observations are plotted together in Figure 1.

The one-step ahead procedure with $\widehat{Y}_{t+1}$ in (3.1) was employed to do prediction. The predicted errors, along with the forecasts given in TYZ (1996) (see last two columns), are reported in Table 1, which shows that the AR model with time-dependent coefficients outperforms the nonlinear

Table 1. The post-sample predictive errors for observations $Y_{101}, \ldots, Y_{109}$

Observation	True Value	Model (3.2)	Error	Model (1.2)	Error
Y_{101}	11.6	10.987	0.613	11.638	-0.038
Y_{102}	11.0	11.311	-0.311	11.609	-0.609
Y_{103}	11.3	11.195	0.105	11.116	0.184
Y_{104}	11.6	11.096	0.504	11.265	0.335
Y_{105}	11.3	11.470	-0.170	11.536	-0.236
Y_{106}	12.1	11.390	0.710	11.344	0.756
Y_{107}	09.2	11.324	-2.124	11.955	-2.755
Y_{108}	10.7	11.141	-0.441	09.727	0.973
Y_{109}	10.4	10.672	-0.271	10.514	-0.114
ASSPE			0.672		1.077

parametric time series model proposed in TYZ (1996) in 6 out of 9 observations (in sense of the absolute error). Furthermore, the average sum of squared prediction errors (ASSPE) is 0.672 which is about 38% less than 1.077 in TYZ (1996). In a sum, model (1.3) does the better performance in prediction due to the flexibility of the AR models with time-dependent coefficients, which have more ability to reduce the bias than the parametric models. Finally, it is worth pointing out that

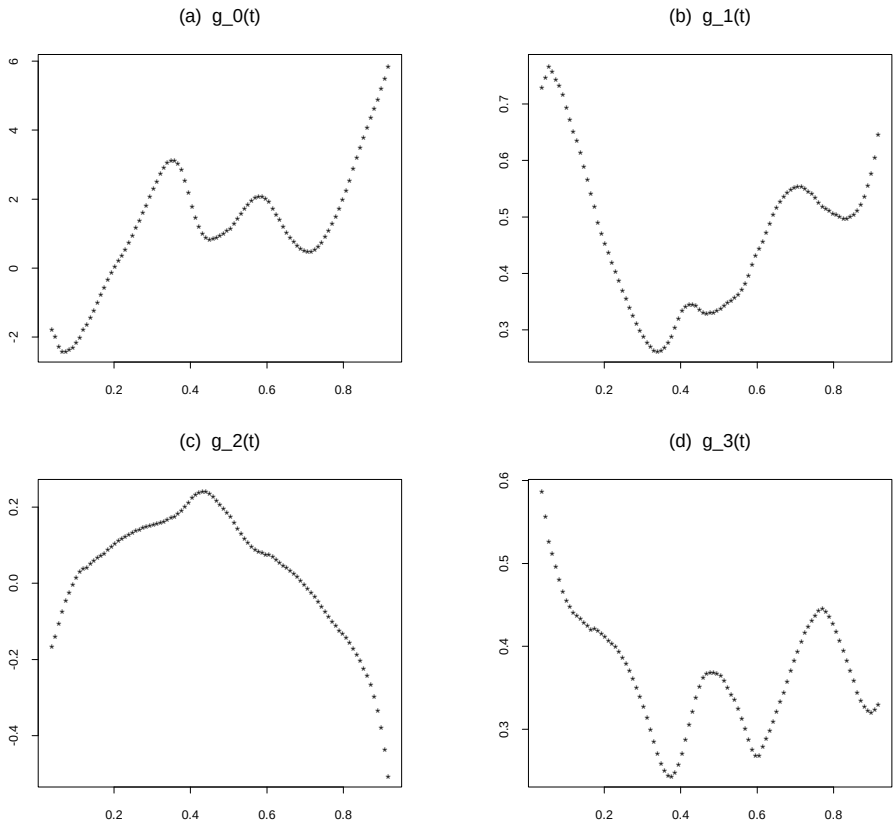

Figure 3: *The scatterplots of the estimated coefficient functions* $g_j(\cdot)(0 \le j \le 3)$. *The* x*-axis is week/109.*

the proposed method is much easier to implement than the TYZ's method that uses the Gibbs sampler.

Next, we applied the goodness-of-fit procedure characterized in Section 2.3 to test whether the linear AR(3) model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \varepsilon_t \tag{3.3}$$

is appropriate for the BOD data. That is to test the null hypothesis $H_0 : g_j(\cdot) = \beta_j (0 \le j \le 3)$. The maximum likelihood estimate of the coefficients in model (3.3) is $(1.1781, 0.5054, 0.0929, 0.2765)$. The test statistic (2.6) is $F = 3.3447$ with degrees of freedom $(15.6426, 78.3574)$. The p-value is 0.0001, which strongly rejects the null hypothesis. This suggests that the autoregressive model with time-dependent coefficients (3.2) gives a much better fit. If the test statistic (2.7) is applied, the test statistic is $f = 6.5573$ with the p-value less than 0.0001. Therefore, both tests give the same conclusion that a simple linear AR model is definitely not appropriate for the BOD data.

Finally, we applied the procedure proposed in Section 2.3 to conduct a test that any lagged variable is significant. Since $\widehat{g}_2(\cdot)$ given in Figure 3(c) seems to fluctuate around x-axis, we may suspect that the variable Y_{t-2} may not be significant in model (3.2). That is we test the hypothesis that the function $g_2(\cdot)$ is zero. The F-test statistic is $F = 3.1877$ with degrees of freedom

(3.3295, 78.3574) and the p-value 0.0241. The normal test statistic is $f = 2.28227$ with the p-value 0.0048. Therefore, the variable Y_{t-2} is statistically significant in model (3.2). Of course, the same procedure can be applied to test whether Y_{t-1} or Y_{t-3} is significant in model (3.2).

4 Concluding Remarks

We proposed procedures for building the varying-coefficient model (2.1) which includes the AR model with time-dependent coefficients. The local linear estimation procedures are developed and implemented, in conjunction with the nonparametric Akaike information criterion, to determine the order of the model. We applied the procedures to the BOD data and obtained an adequate model. The analysis performed using the Akaike information criterion showed that our model outperforms TYZ's model. This leads us to conclude that the varying-coefficient model (2.1) is simple and easy to be used to fit certain time series data by employing the methodology developed here.

References

Brockwell, R.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd. Springer: New York.

Cai, Z., Fan, J. and Yao, Q. (1998). Functional-coefficient regression models for nonlinear time series. Submitted to *Journal of the American Statistical Association*, revised.

Cai, Z. and Tiwari, R.C. (1998). Nonparametric modeling of varying-coefficient autoregressive models. *Technical Report No.98-CT1*, Department of Mathematics, University of North Carolina at Charlotte.

Chen, R. and Tsay, R.S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, **88**, 298-308.

Cleveland, W.S. and Devlin, S. (1988). Locally-weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

Fuller, F.C., Jr. and Tsokos, C.P. (1971). Time series analysis of water pollution data. *Biometrics*, **27**, 1027–1034.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hastie, T.J. and Tibshirani, R. J. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B*, **55**, 757-796.

Hurvich, C.M., Simonoff, J.S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, **60**, 271-293.

Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

Padgett, W.J. and Rao, A.N.V. (1979). Estimation of BOD and DO probability distribution. Proc Paper 14620, *Journal of Environmental Engineering Division ASCE*, **105**(EE3), 525–533.

Papadopoulos, A.S. and Tiwari, R.C. (1994). Bayesian approach for BOD and DO when the discharged pollutants are random. *Ecological Modeling*, **71**, 245–257.

Papadopoulos, A.S., Tiwari, R.C., and Muha, M.J. (1991). Bootstrap procedures for time series analysis of BOD data. *Ecological Modeling*, **55**, 57–65.

Tiwari, R.C. and Dienes, T.P. (1994). The Kalman filter model and Bayesian outlier detection for time series analysis of BOD data. *Ecological Modeling*, **73**, 159–165.

Tiwari, R.C., Yang, Z. and Zalkikar, J.N. (1996). Time series analysis of BOD data using the Gibbs sampler. *Environmetrics*, **7**, 567–578.

Tong, H. (1990). *Nonlinear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford.