Stoch Environ Res Risk Assess (2010) 24:19–28
DOI 10.1007/s00477-008-0295-6

ORIGINAL PAPER

Tree-based threshold modeling for short-term forecast of daily maximum ozone level

Sung Eun Kim

Published online: 6 December 2008
© Springer-Verlag 2008

Abstract This paper proposes a simple class of threshold autoregressive model for purpose of forecasting daily maximum ozone concentrations in Southern California. Linear time series model has been widely considered in environmental modeling. However, this class of models fails to capture the nonlinearity in ozone process and the complexity of meteorological interactions with ozone. In this article, we used the threshold autoregressive models with two classes of regimes; periodic and meteorological regimes. Days in week were used for the periodic regimes and the regression tree method was used to define the regimes as a function of meteorological variables. As the reference model we used the autoregressive model with lagged ozone and various lagged meteorological variables as the covariates. The proposed models were applied to a 3-year dataset of daily maximum ozone concentrations obtained from five monitoring stations in San Bernardino County, CA and their forecast performances were evaluated using an independent year-long dataset from the same stations. The results showed that the threshold models well capture the nonlinearity in ozone process and remove the nonstationarity in model residuals. The threshold models outperformed the non-threshold autoregressive models in day-ahead forecasts. The tree-based model showed slightly better performance than the periodic threshold model.

Keywords Ozone forecast · Nonlinear time series · Meteorological adjustment · Autoregression

S. E. Kim (✉)
Department of Mathematics and Statistics, California State University Long Beach, Long Beach, CA 90840, USA
e-mail: skim43@csulb.edu

1 Introduction

Ground-level ozone (O_3), a major element of urban smog, is the one of the most complex, difficult to control, and pervasive pollutants. Ozone can be produced by photochemical reactions between primary pollutions such as oxides of nitrogen (NO_x) and volatile organic compounds (VOC) in the presence of sunlight. Some of the major sources of NO_x and VOC are emissions from industrial facilities and electric utilities, motor vehicle exhaust, gasoline vapors, and chemical solvents. Ozone concentrations can reach unhealthful level when the weather is hot and sunny with little or no wind. Despite decades of effort in reducing air pollution, California suffers from the worst air quality in the nation and over 30 million individuals in the state are exposed to unhealthful levels of ozone. Southern California area currently has the highest ozone level in the US. The area's sunny climate and mountainous topography accelerates the formation of ozone episodes. The meteorological adjustment of ozone in various statistical modeling has been considered in a number of articles (e.g. Fiester and Balzer 1991; Bloomfield et al. 1996; Davis and Deistler 1998). A comprehensive review can be found in Thomson et al. (2001). We consider daily ambient and soil temperature, global radiation, relative humidity, evapotranspiration, and wind components as predictors in the statistical forecast model proposed in this article.

One of the most important roles of air pollution forecasts is to provide the public early warnings of high pollution levels. For this, the ability of a forecast model to forecast pollution levels accurately at least one day ahead is important. The purpose of this study is to develop and to implement operational statistical model to forecast daily maximum ozone levels. Of particular interest is to model the nonlinear dynamics of ozone processes using linear approximation.

Linear time series model, like autoregressive model, has been widely considered in environmental modeling (e.g. Robeson and Steyn 1990; Chaloulakou et al. 1999). However, this class of model fails to capture the complexity of meteorological interactions and nonlinearity in ozone process (e.g. Kim and Kumar 2005). For this a nonlinear model can be considered (e.g. Bloomfield et al. 1996). However, nonlinear modeling is rather complex with too many possible structures and is not suitable for multi-step forecasts. In this paper, we consider the threshold model of Tong (1983, 1990). The threshold model is considered as separates linear time series models where the separations are made in terms of regimes where the dynamic of the ozone system changes. For instance, temperature correlation of ozone measurements at high ozone level and at low ozone level may well be different. Two classes of regime are considered. We first naturally consider day in week to define the threshold regimes so that ozone data can be modeled linearly within each of 7 days in week. We also define regimes as a function of meteorological variables and lagged ozone concentrations. A regression tree approach is used to recursively partition data to meteorologically homogeneous subsets or clusters. We then fit separate linear models to observations in each cluster. For the purpose of forecasting, we use a global linear model with indicator variables.

The data are described in Sect. 2 and a brief outline of the regression tree method is given in Sect. 3. The model classes are presented in Sect. 4. Section 5 gives the test statistics for nonlinearity and the model evaluation parameters. Section 6 presents an empirical application of the proposed models to the Southern California ozone data. Model identification, comparison, and the performance of one-day-ahead forecast are also given in Sect. 6. Concluding remarks are given in Sect. 7.

2 Data

The South Coast Air Basin, which covers Los Angeles, Orange, Riverside, and San Bernardino counties in Southern California, currently is not in compliance with National Ambient Air Quality Standard (NAAQS) for ozone. According to the American Lung Association, the four counties graded “F” for ozone and San Bernardino ranked as the most ozone-polluted county in the US. Riverside and Los Angeles ranked the third and the fourth, respectively. The California Air Resources Board (ARB) gathers air quality data for the State of California and makes it available to public. Data used in modeling fitting are based on daily maximum ozone records (in ppb) from five monitoring stations in San Bernardino County from January 2001 to December 2003. The stations are listed in Table 1. For the purpose of evaluating the performance of short term ozone forecast, 2004 data from the same station is used. Thus, the evaluation uses an independent data set. The meteorological data used in this study came from the University of California Statewide Integrated Pest Management Program (UC IPM: http://www.ipm.ucdavis.edu). It consists of daily ambient and soil temperature, global radiation, relative humidity, evapotranspiration, and west–east and south–north wind components (see Table 2).

Table 1 Ozone monitoring stations in San Bernardino County, CA

Stations	AIRS ID	Latitude	Longitude	Missing % (01–04)
Crestline	060710005	34.24	−117.28	2
Fontana	060712002	34.10	−117.49	4
Redlands	060714003	34.06	−117.15	0
SB-4th St	060719004	34.11	−117.27	1
Upland	060711004	34.10	−117.63	0

Table 2 Meterological variables

1	Temp	Daily maximum air temperature (F)
2	Eto	Daily reference evapotranspiration (in.)
3	Prec	Daily total precipitation (in.)
4	Rh	Daily maximum relative humidity (%)
5	Soil	Daily maximum soil temperature (F)
6	Solar	Daily global radiation (Watts/m^2)
7	Wind.u	Daily west−east wind component (m/s)
8	Wind.v	Daily south–north wind component (m/s)

3 Outline of regression trees

Regression tree model recursively partitions data into meteorologically homogeneous subsets toward the response and is very useful when data shows nonlinear relation between response and various predictors. Regression tree method has been used in environmental studies (e.g. Burrows et al. 1995; Huang and Smith 1999). Detail discussion of regression tree can be found in Breiman et al. (1984).

The tree method consists of two sequential processes; splitting and pruning. Staring with a global node having all observations in the node, the splitting process partitions the data in the current node (parent node) into two parts (child nodes) with maximum homogeneity. Under the normal assumption, the deviance for node k is defined as

$$D_k = \sum_i (y_i - \hat{\mu}_k)^2,$$

where $\hat{\mu}_k$ is the sample mean of the observed responses y_i's in the given node. Then the parent node is split into two children nodes (left and right) which minimize the difference between the deviance of the parent node and the sum deviance of the children nodes. Then for each split node this splitting process is applied. This procedure continues until the number of observations in the node is less than a pre-specified limit (usually 10% of the training sample size) or the difference between the deviance of the parent node and the sum deviance of the children nodes is small enough (usually less than 1% of the parent deviance).

The grown tree from the splitting process is usually very complex and causes over-fitting problems when it is applied to an independent data for prediction. The pruning process chooses a right size of tree by cutting off insignificant nodes and subtrees. Breiman et al. (1984) first introduced the minimum cost-complexity pruning method. The tree is pruned to minimize the cost-complexity factor,

$$R_\alpha(T) = R(T) + \alpha\,\mathrm{size}(T),$$

where $R(T)$ is the mean square error of the predictions of the subtree T, α the cost-complexity parameter and size(T) the number of nodes of T. The pruning algorithm used in this study uses a tenfold cross-validation to compute a cost-complexity factor. The nodes in the final tree serve as regimes in the threshold model. We used the S-Plus library *rpart* which is available from Stat Lib (http://lib.stat.cmu.edu/S/). Detailed discussion on *rpart* is available in Venables and Ripley (2002).

4 Statistical model

4.1 Reference model

We consider an autoregressive model with an exogenous vector process (ARX) as the reference model for this study. The daily observed time series $y(t)$ with mean μ can be modeled as a regression form:

$$\phi(B)y(t) = \mu + \boldsymbol{\beta}'\mathbf{x}(t) + w(t), \qquad t = q+1,\ldots,n \tag{1}$$

where B is the backshift operator defined as $B^k y(t) = y(t-k)$, $k = 0, \pm1, \pm2,\ldots$, and $\phi(B)$ is polynomials given by $\phi(B) = 1 - \sum_{j=1}^{p}\phi_j B^j$ with $\phi_p \neq 0$ where p, the order of the autoregressive model, is a positive integer. $w(t)$ is assumed to be zero mean white noise process. The exogenous vector $\mathbf{x}(t) = (\mathbf{x}_1(t),\ldots,\mathbf{x}_l(t))'$ contains l exogenous processes with lags $r_1,r_2,\ldots,r_l$, respectively, $\mathbf{x}_k(t) = (x(t),x(t-1),\ldots,x(t-r_k))$, $k = 1,2,\ldots,l$ and $q = \max\,(p,r_1,\ldots,r_l)$. The order $r_1,r_2,\ldots,r_l$ may involve seasonal components. The parameter vector $\boldsymbol{\beta}$ are denoted by $\boldsymbol{\beta} = (\gamma_1, \gamma_2,\ldots, \gamma_l)'$, where $\gamma_k = (\gamma_{k0}, \gamma_{k1},\ldots,\gamma_{kr_k})$, $k = 1,\ldots, l$. For the exogenous processes, we used the meteorological series listed in Table 2. Two temporal variables which may affect the baseline measurements of ozone level are added as a binary signals defined by

$$\mathrm{week}(t) = I(t = \text{weekend day}) \text{ and}$$
$$\mathrm{season}(t) = I(t = \text{ozone season day}),$$

where $I(A)$ is the indicator function equal to one if the event A occurs and zero otherwise. Weekend days are Saturdays and Sundays and ozone season days are days between April 1 and October 31. The model can be written as a multiple regression form in which the vector of predictor is

$$\begin{aligned}\mathbf{z}(t) &= (1, z_1(t), z_2(t),\ldots, z_M(t))' \\ &= (1, y(t-1),\ldots, y(t-p), \mathbf{x}_1(t),\ldots,\mathbf{x}_l(t))',\end{aligned} \tag{2}$$

with the corresponding parameter vector of

$$\boldsymbol{\theta} = (\theta_0, \theta_1,\ldots,\theta_M)' = (\mu, \phi_1,\ldots,\phi_p,\ \boldsymbol{\beta})', \tag{3}$$

where M denotes the number of predictors.

For a comparison purpose we also consider the model without exogenous processes (AR model). The order p can be estimated using the iterative procedure of Box and Jenkins (1976) and the order $r_1,r_2,\ldots,r_l$ can be obtained from the cross-correlation functions (CCF) between the pre-whitten response and each of the pre-whitten exogenous processes (see Box and Jenkins 1976 for detailed discussions). Some alternative optimality criteria, e.g., Akaike's AIC (Akaike 1974) and Schwarz's SIC (Schwarz 1978), can be also used. However, in order to avoid overfitting problems, the procedure by Box and Jenkins has been suggested in air pollution modeling (Millionis and Davies 1994).

4.2 Threshold autoregressive model with an exogenous vector process (TARX)

Due to the nonlinear feature of an ozone process, the ARX model, even with use of high dimensional predictor vectors, shows seasonal nonstationarity in residuals that cause systematic over- and under-predictions (see Bauer et al. 2001; Fasso and Negri 2002; Kim and Kumar 2005). Since first introduced by Tong (1983), the threshold autoregressive model has been an important tool for modeling nonlinear phenomena in many areas, including environmetrics (e.g. Bauer et al. 2001; Fasso and Negri 2002; Kim and Kumar 2005) and economics (e.g. Tiao and Tsay 1994; Potter 1995; Hansen 1999).

The threshold ARX model (TARX) separates the autoregressive model in terms of *regimes* such that ozone can be modeled linearly within each regime. The threshold model is considered as separated ARX models where the separations are made in term of regimes where the dynamic of the system changes:

$$\begin{aligned}y(t) &= \boldsymbol{\theta}'\mathbf{z}(t) + \sum_{k}^{K-1}\boldsymbol{\theta}_k'\mathbf{z}(t)\delta_k(t) + w(t), \\ \delta_k(t) &= \mathrm{I}(\boldsymbol{\tau}'\mathbf{z}(t) \in R_k)\end{aligned} \tag{4}$$

where, $\mathbf{z}(t)$ and $\boldsymbol{\theta}$ are as in (2) and (3) and the *regimes*, R_k, $k = 1,\ldots, K$, are mutually exclusive and exhaustive regions in real line, R. $\boldsymbol{\tau}$ is a known column vector of size n and $\mathbf{I}(\cdot)$ is the indicator function. The threshold model is flexible and very effective when regimes are well defined such that ozone can be modeled linearly within regimes. This paper uses and tests two classes of regimes; periodic and meteorological regimes. We use day in week, $h(t) = 1,2,\ldots,7$, as the periodic regimes. For this regimes $\delta_k(t)$ takes one if $h(t) = k$, $k \in \{1,2,\ldots,7\}$ and zero otherwise. The TARX model with such regimes is called periodic TARX model (P-TARX). To reduce the number of regimes, regimes with similar ozone dynamics can be combined. To identify the proper number of regimes model evaluation parameters in Sect. 5 are used. Regimes can be also defined as a function of meteorological variables and lagged ozone concentrations through various classification approaches, like clustering and tree model. We use the regression tree for this application and call such model tree-based TARX model (T-TARX). Once regimes are identified, we fit the ARX model for each regime separately. However, for prediction purpose, we fit the ARX model with indicator function given in (4).

5 Hypothesis test and model evaluation parameters

5.1 Testing hypothesis

The hypothesis for testing varying coefficients over regimes can be formulated as

$$H_0 : \boldsymbol{\theta}_k = \mathbf{0} \quad \text{for all } k \in \{1, 2, \ldots, K-1\} \text{ and } k \neq j,$$

where K is the number of regimes.

For given $q = \max\,(p,r_1,\ldots,r_l)$, the effective number of observations in the regression in (4) is $n - q$. The number of parameter is $K(M + 1)$ for the full model (4) and $M + 1$ for the reduced model under H_0. Then, when the model under H_0 is correct, the general linear test statistic

$$F = \left(\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}\right)\left(\frac{n - q - K(M+1)}{(K-1)(M+1)}\right) \tag{5}$$

has a central F-distribution with $(K - 1)(M + 1)$ and $n - q - K(M + 1)$ degrees of freedom. Here, RSS_0 and RSS_1 are the residual sum of squares under the reduced model and the full model, respectively. Once the test rejects the null hypothesis, it indicates that not all regression coefficients are constant and thus the threshold regression is suggested. For each fitted model, we also consider two versions of the portmanteau test of goodness of fit. Let $\hat{\rho}_k^2$ be the kth sample autocorrelation coefficient of the residuals from the fitted model, then in the case of white residuals, Ljung and Box (1978) showed that the statistic

$$Q_{\mathrm{LB}} = n'(n'+2)\sum_{k=1}^{h}(n'-k)^{-1}\hat{\rho}_k^2 \tag{6}$$

has an approximate χ^2 distribution with h degrees of freedom. Here n' denotes the number of sample used to calculate $\hat{\rho}_k^2$. Monti (1994) proposed a similar statistic which replaces the autocorrelation $\hat{\rho}_k^2$ with the partial autocorrelation.

5.2 Model evaluations parameters

Several model evaluation parameters widely used in environmental study are considered. The consideration is based on their use in air pollution model evaluation studies. If we denote y_t the observed values, $\hat{y}_t$ the predicted values, $\bar{y}$ the sample mean of observed values, $\bar{\hat{y}}$ the sample mean of predicted values, and p the number of parameters in the model, one can compute:

Root mean square error $(RMSE)$:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_t (y(t) - \hat{y}(t))^2}{n-p}} \tag{7}$$

Coefficient of determination (R^2):

$$R^2 = \frac{\sum_t (\hat{y}(t) - \bar{y})^2}{\sum_t (y(t) - \bar{y})^2} \tag{8}$$

RMSE is an unbiased estimator of the regression error variance and most commonly used statistic in cross-validation schemes. It takes the number of parameters p into account through $n - p$ in the denominator. R^2 is the proportionate reduction of total variation in the time series associated with the use of the model. Three parameters below are also introduced to assess the forecast performance of air quality models (e.g. Kumar et al. 1999).

Fraction bias (FB): $$\mathrm{FB} = \frac{2(\bar{y} - \bar{\hat{y}})}{(\bar{y} + \bar{\hat{y}})} \tag{9}$$

Normalized mean square error $(NMSE)$:

$$\mathrm{NMSE} = \frac{(\bar{y} - \bar{\hat{y}})^2}{\bar{y}\bar{\hat{y}}} \tag{10}$$

Factor of two $(Fa2)$:

$$\mathrm{Fa2} = \text{fraction of data which satisfy } 0.5 \le \frac{\hat{y}(t)}{y(t)} \le 2.0 \tag{11}$$

FB is the normalized mean bias varying between −2 and +2 and has a value of zero for an ideal model. NMSE emphasizes the scatterness of residuals in the entire data set. The normalization by the product in the denominator assures that the statistic will not be biased towards over- or under-predictions. Smaller values of NMSE denote better model performance. Fa2 is defined as the percentage of the predictions within a factor of two of the observed values. The ideal value for Fa2 should be 1. In addition, the percentage of the predicted values within ±5 and ±10 ppb of the observed values are calculated to get an idea of the forecasting ability of the models.

6 Ozone prediction in San Bernardino, CA

Periodic threshold ARX (P-TARX) and tree-based threshold ARX (T-TARX) models are fitted for 3 years of training data (2001–2003) from each monitoring stations in San Bernardino County, CA. Then, the fitted models are applied to 2004 data to evaluation the performance of day-ahead forecasts.

6.1 Model identification

To identify an initial ARX model, the iterative procedure by Box and Jenkins (1976) is applied to each of training data. The partial autocorrelation function of ozone shows one dominant peak at lag one and this suggests AR(1) for an initial model. Further investigations of the cross-correlation functions (CCF) between the pre-whitten response and each of the eight pre-whitten exogenous processes show dominant peaks at lag 0 or lag 0 and 1. We also include the two indicator variables, week(t) and season(t), in the model. This defines the reference ARX model. A stepwise model selection procedure is applied to fit the final ARX model.

Table 3 contains the results from fitting the AR and the ARX models for each station. It is clear that the ARX model fits significantly better than AR; R^2 is increased by 5–16% and RMSE is decreased by 13–24%. This shows the importance of meteorological variables in ozone modeling. The estimated values of the AR and the final ARX parameters for the Upland data are listed in Table 4. Figure 1 draws the P values of the portmanteau χ^2 tests with the lag up to 30. This shows that both AR and ARX models reject the null hypothesis of white errors. The autocorrelation function (ACF) plots (Fig. 2a, b) also pronounce the weekly behaviors of the model residuals. This is due to the nonlinear feature of ozone process and this motivates the use of a threshold model.

To account for the nonlinearity in the model, the threshold models in Sect. 4.2 are applied to the ARX model. The weekly behavior of the model residuals suggests the use of the day of week as periodic regimes (P-TARX). The F test statistics (5) range from 3.8 to 4.1 with P values all less than 0.001. This rejects the null hypothesis of constant coefficients and supports the use of the threshold model. Figure 3 gives the least squared estimates of the coefficients with 95% confidence limits under the P-TARX model for each of 7 days of week. The plot well discloses the changes of the regression coefficients over the regimes. The results from the fitted the P-TARX model are summarized in Tables 3 and 4. To avoid singularity which may occur during the estimation procedure, the estimated coefficient for *week* variable is assumed to be the same across regimes. A stepwise modeling selection procedure is applied to find the final model. Comparing to the results from the ARX model, the P-RARX model increases R^2 by about 3% and decreases RMSE by 3–17%.

We also consider regimes defined as a function of meteorological variables through the regression tree (T-TARX). Figure 4 shows examples of the nonlinear relations between ozone and meteorological variables for Upland data. Figure 4a indicates that the data can be partitioned into two regimes with a threshold at around temp(t) = 85. Similarly, the data can be partitioned with a threshold at around eto(t) = 0.14. The regression tree for Upland data is shown in Fig. 5 containing six clusters, indexed 1 through 6 from the left. Cluster 6 has the highest average daily maximum ozone level at 114 ppm containing 73 days. The corresponding conditions for the cluster are $y(t-1) > 55$, temp(t) > 88.2, and week(t) = *1*. Cluster 1 contains 436 day and has the lowest average daily maximum with the conditions of $y(t-1) < 55.5$ and eto(t) < 0.135. The six clusters are used as threshold regimes in the T-TARX for Upland data. Separate tree model is applied to each of the five data sets. The numbers of cluster range from five to seven. The F test statistics (5) range from 5.2 to 6.1 with P values all less than 0.001.

Table 3 Comparison of fitted models

Model	Crestline		Fontana		Redlands		SB-4th St		Upland	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
AR	0.74	15.32	0.66	19.45	0.70	17.51	0.68	17.22	0.64	18.05
ARX	0.80	13.37	0.80	14.89	0.81	13.96	0.81	13.11	0.80	14.89
P-TARX	0.82	12.92	0.83	13.91	0.83	13.41	0.84	12.28	0.83	12.37
T-TARX	0.83	12.56	0.85	12.25	0.84	13.10	0.86	11.60	0.85	11.89

Table 4 Model identification for Upland data

	Const	Oz (t)	Temp (t)	Temp ($t-1$)	Soil (t)	Soil ($t-1$)	Eto (t)	Pr (t)	Week (t)
AR									
	11.13 (1.2)	0.80 (0.02)	na	na	na	na	na	na	na
ARX									
	−36.92 (4.2)	0.45 (0.02)	0.6 (0.09)	−0.59 (0.08)	2.04 (0.41)	−1.34 (0.40)	103.3 (10.3)	2.45 (1.47)	13.86 (0.90)
P-TARX									
1	−60.30 (10.9)	0.61 (0.06)	1.24 (0.21)	−0.88 (0.22)	4.35 (1.09)	−3.39 (1.05)	–	–	–
2	–	0.53 (0.04)	0.72 (0.18)	−0.69 (0.18)	–	–	83.0 (21.2)	8.13 (2.90)	–
3	–	0.48 (0.05)	–	–	–	–	163.4 (17.2)	–	–
4	−26.12 (9.9)	0.58 (0.07)	0.74 (0.22)	−0.80 (0.19)	0.62 (0.22)	–	77.9 (25.5)	–	–
5	−34.83 (8.8)	0.27 (0.06)	–	–	4.07 (0.86)	−3.28 (0.85)	132.0 (20.4)	–	–
6	−14.21 (7.4)	0.60 (0.06)	0.95 (0.20)	−0.73 (0.19)	–	–	111.5 (23.7)	7.56 (3.17)	–
7	−51.66 (12.1)	0.67 (0.08)	1.19 (0.26)	−0.92 (0.24)	0.72 (0.24)	–	83.6 (29.6)	–	–
T-TARX									
1	–	0.35 (0.04)	0.16 (0.09)	−0.27 (0.07)	1.32 (0.38)	−0.62 (0.37)	123.2 (12.3)	–	–
2	−43.87 (7.8)	0.39 (0.10)	0.65 (0.15)	−0.56 (0.14)	0.88 (0.14)	–	73.2 (26.5)	–	–
3	–	0.27 (0.06)	0.88 (0.23)	−0.64 (0.20)	–	–	123.6 (25.3)	–	–
4	−69.62 (30.7)	0.42 (0.08)	1.33 (0.40)	–	13.95 (2.29)	−13.50 (2.06)	−92.7 (47.0)	–	–
5	–	0.38 (0.06)	0.77 (0.40)	−1.09 (0.34)	3.91 (1.67)	−3.22 (1.63)	111.3 (35.3)	–	–
6	–	0.51 (0.10)	–	−0.88 (0.50)	1.90 (0.59)	–	–	–	–

Values in parentheses are the standard errors for the estimates

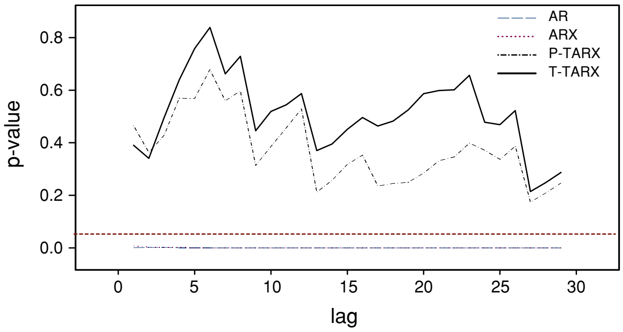


Fig. 1 P values of Portmanteau χ^2 test

Table 3 shows that the T-TARX model fits the data slightly better than the P-TARX. The final fitted model for Upland data is in Table 4.

Both P-TARX and T-TARX model remove the autocorrelation in residuals (Fig. 2c, d). The portmanteau tests for Upland data in Fig. 1 support this result with the P values ranging from 0.2 to 0.8. Data from other stations show similar results.

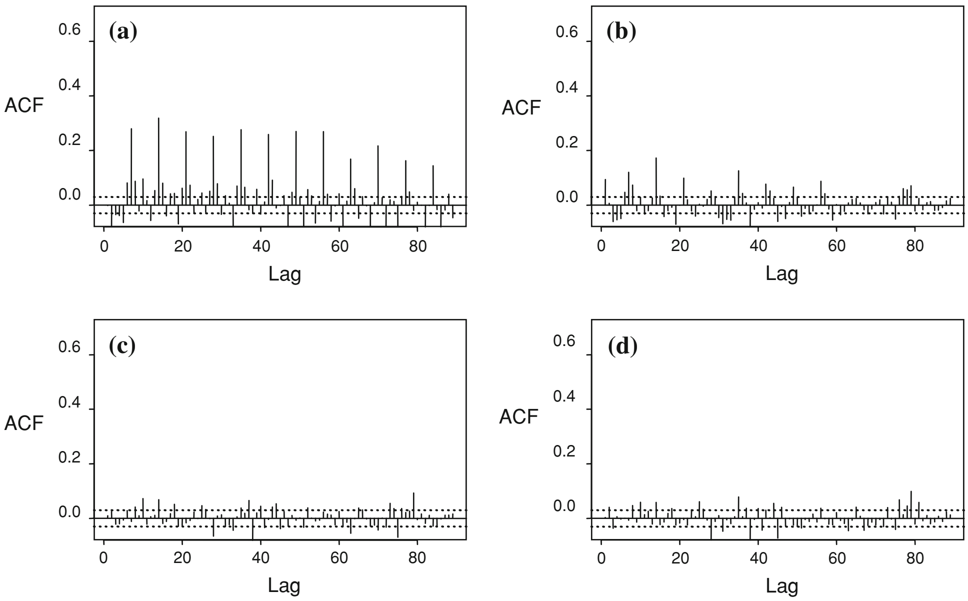
6.2 Forecasting

The performance of the forecasting models was evaluated using the five parameters in Sect. 5.2. To avoid further complications, it is assumed that meteorological processes are observable or at least predictable from a weather forecasting system. In this sense, forecasting procedures discussed in this section is conditional to known exogenous processes. Since similar results are found for data from all five stations, we only discuss the results for Upland data in this section.

Table 5 summarizes the evaluation of one-day-ahead forecast performances. All models have FB and NMSE values close to zero showing that they are acceptable in terms of unbiasness and scatterness of the mean residuals. Both P-TARX and T-TARX models clearly outperform the AR and the ARX model. As regard the two threshold models, the T-TARX model is slightly better in terms of all evaluation parameters.

Figure 6 draws the daily profile of one-day-ahead forecasts of daily maximum ozone level from different models for July to August, 2004 at Upland station. During the months of July and August, the study area usually shows high daily maximum ozone levels. Both P-TARX and T-TARX models give similar results, so the daily forecasts from the P-TARX model is omitted from the plot. The threshold model clearly outperforms the constant coefficient models and seems to give a satisfactory forecasting performance. Both T-TARX and ARX model nicely capture the temporal pattern of the ozone maxima, however, the ARX model under predicts the peaks at over 100 ppb. The forecast performance for the AR model is not satisfactory; the forecasts are consistently shifted by a day and this is due to the heavy dependence on the immediate past ozone readings in the model. The T-TARX models appear

Fig. 2 ACF of residuals from fitting various models at Upland station



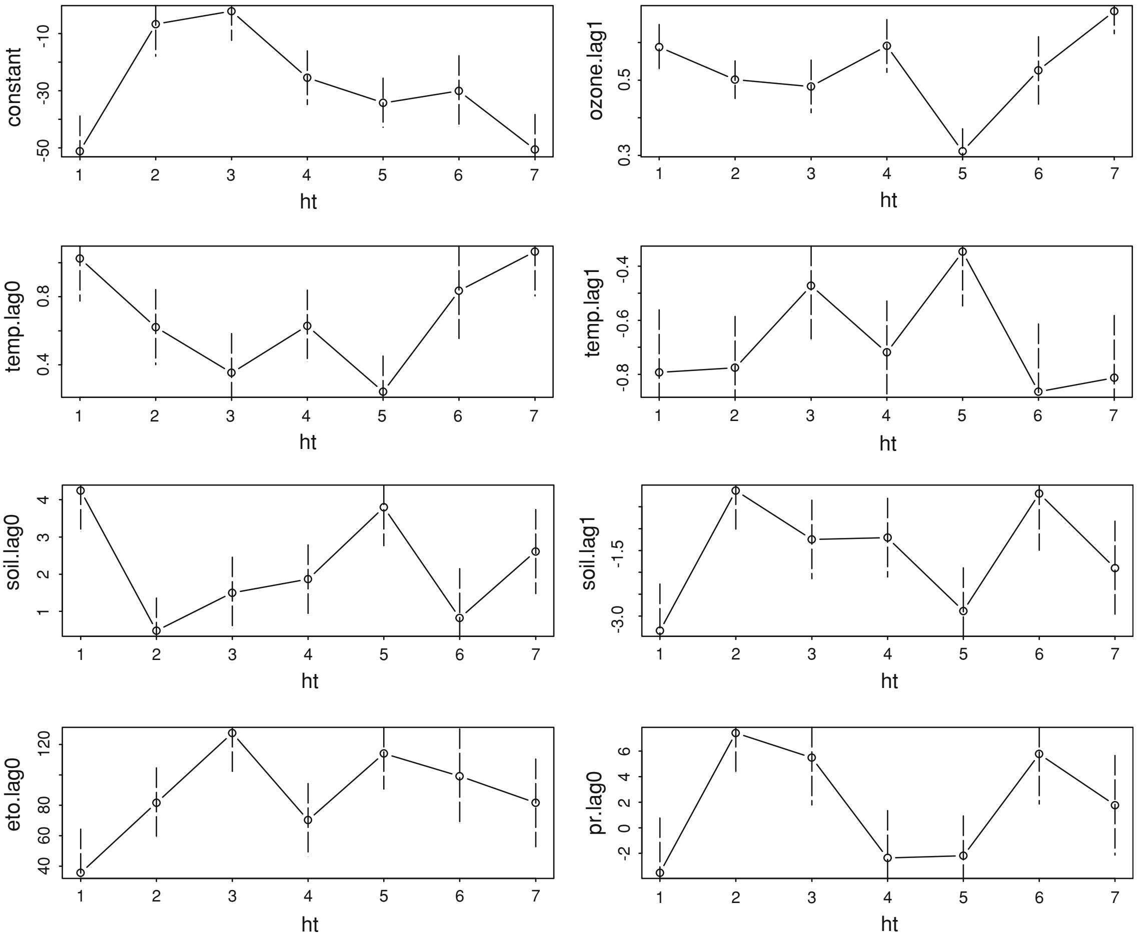


Fig. 3 Coefficient estimates from P-TARX model with 95% confidence limits

to agree very well with the observed values on most days. Figure 6b shows that the absolute forecast errors in each model. For 263 of 366 days (72%) the forecasted values from the T-TARX are within ±10 ppb of the actual values. The percentage is 54% for the AR and 64% for the ARX model.

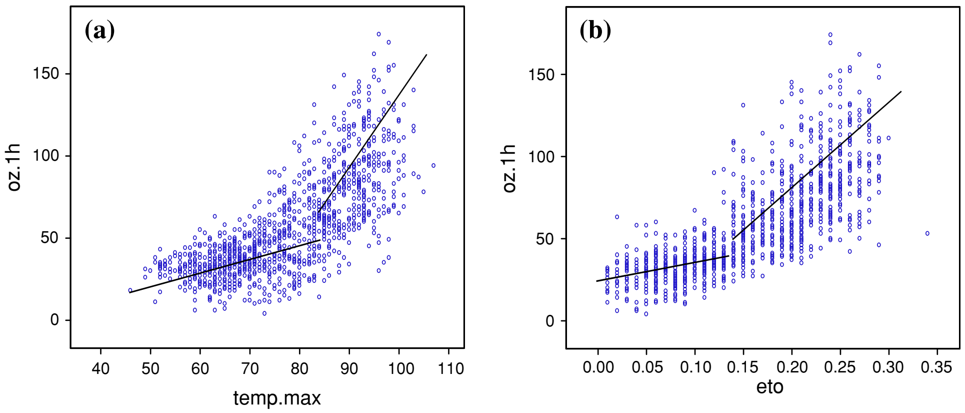


Fig. 4 Nonlinear relation between daily maximum ozone and meteorological variables; (**a**) daily maximum temperature and (**b**) daily reference evapotranspiration

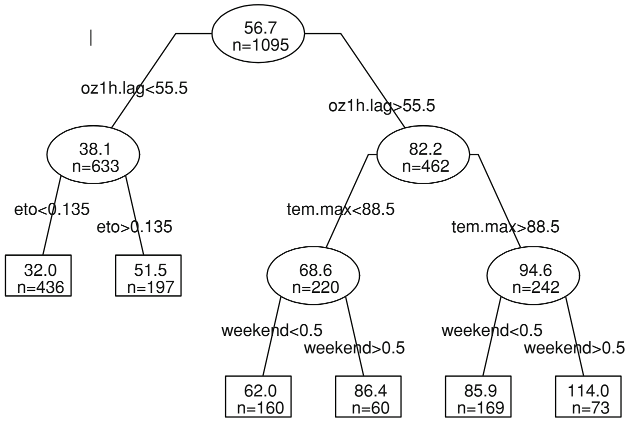


Fig. 5 Regression tree for Upland data

Table 5 Model evaluation parameters for day-ahead forecasts of 2004 daily maximum ozone level at San Bernadino, CA for different model classes

	AR	ARX	P-TARX	T-TARX
R^2	0.63	0.78	0.81	0.83
RMSE	15.81	12.16	11.58	11.21
FB	−0.07	−0.02	−0.02	−0.02
NMSE	<0.01	<0.01	<0.01	<0.01
Fa2	0.973	0.980	0.981	0.990
±5	0.30	0.34	0.36	0.43
±10	0.54	0.64	0.67	0.72

7 Conclusion

Daily maximum ozone data from five monitoring stations in San Bernardino County, CA have been used to develop a statistical forecasting model. The autoregressive model with exogenous vector process (ARX) is used as the reference model. The lagged meteorological variables are used as the exogenous variables. Due to the nonlinearity in ozone data, the ARX model fails to remove the nonstationarity in errors and is not suitable for ozone prediction. In this paper, high-resolution threshold autoregressive models (TARX) with two classes of regimes are proposed; periodic and meteorological regimes. The periodic threshold model makes use of the day in week as the threshold regimes. This model equals to separate ARX model for each day in week. We also considered the tree-based TARX model where regimes are defined as a function of meteorological variables through the regression tree model. Application of the proposed model to three-year training data in San Bernardino indicates that both threshold models successfully removed the nonstationarity in model residuals and the RMSE was decreased by up to 17%. The fitted models are applied to new year-long data from the same stations and the performance of the day-ahead forecast was evaluated using five parameters. The result indicates that the threshold models clearly outperform the non-threshold model in all five model evaluation parameters. The tree-based TARX model showed slightly better performance than the periodic TARX model.

By removing the nonstationarity in model residuals, the TARX models well predict the days with high peaks of ozone concentrations. This would make it possible to use the predicted ozone concentrations to predict air quality index (AQI). Finally, we note that, the autocorrelation in time series in each tree cluster has been handled using ARX model. However, one may apply an appropriate transformation prior to the tree construction to remove the autocorrelation in time series and to have the deviance used in tree construction more meaningful. The statistical assessment of this alternative requires further investigation.

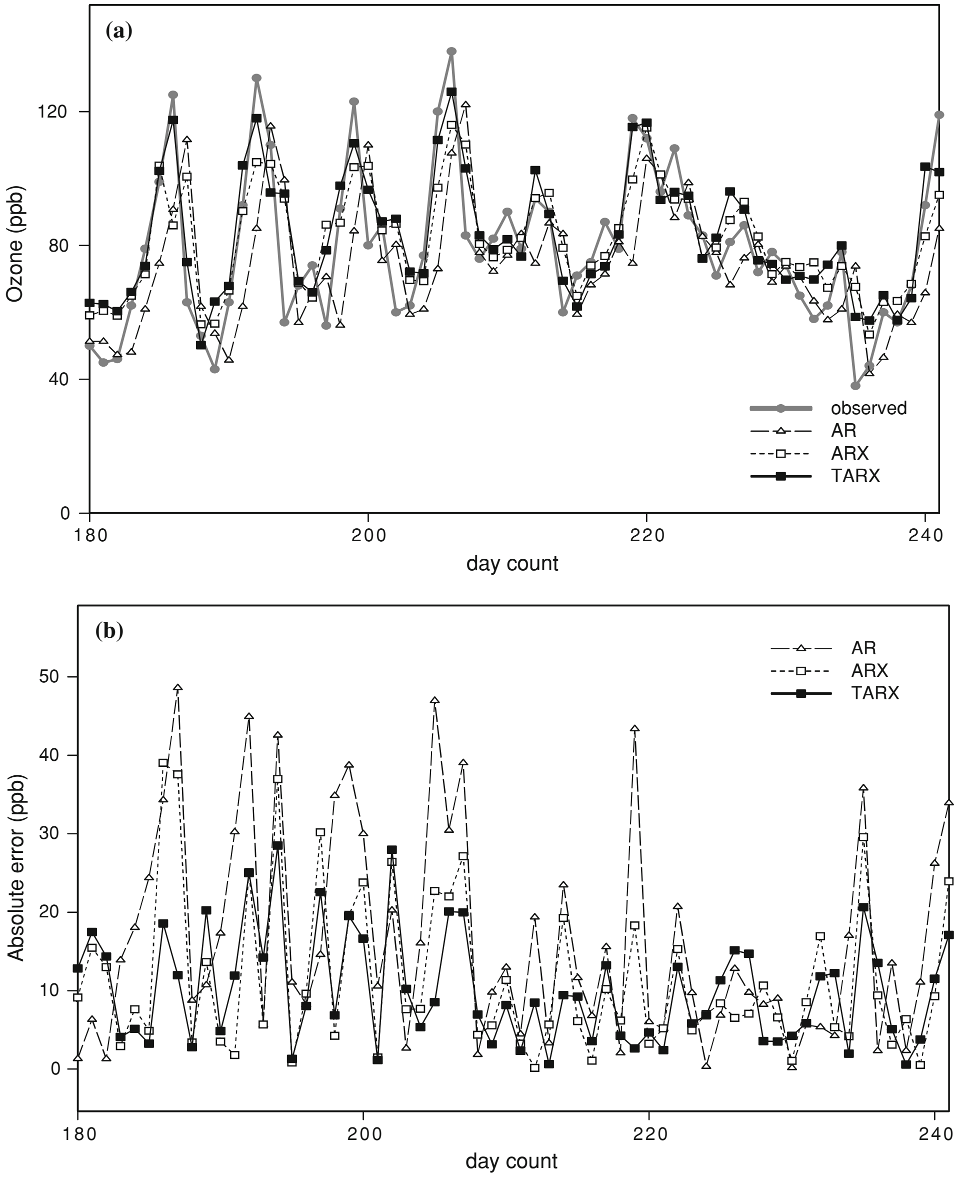


Fig. 6 Performance of day-ahead forecasts of daily maximum ozone level for July to August, 2004 at Upland station using various models; (**a**) actual observation and forecasts, (**b**) absolute errors

Acknowledgment This work was supported in part by a grant from California State University, Long Beach, SCAC program.

References

Akaike H (1974) A new look at statistical model identification. IEEE Trans Automat Contr AC-19:716–723

Bauer G, Deistler M, Scherrer W (2001) Time series models for short term forecasting of ozone in the eastern part of Austria. Environmetrics 12:117–130

Bloomfield P, Royle JA, Steinberg LJ, Yang Q (1996) Accounting for meteorological effects in measuring urban ozone levels and trends. Atmos Environ 30(17):3067–3077

Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control, revised edn. Holden Day, San Francisco

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall, London

Burrows WR, Benjamin M, Beauchamp S (1995) CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic Regions of Canada. J Appl Meteorol 34:1848–1862

Chaloulakou A, Assimacopoulos D, Lekkas T (1999) Forecasting daily maximum ozone concentrations in the Athens Basin. Evrion Monit Assess 56:97–112

Davis MD, Deistler M (1998) Modeling ozone in the Chicago urban area. In: Nychka D, Piegorsch W, Cox LH (eds) Case studies in environmental statistics. Lecture notes in statistics. Springer, Berlin

Fasso A, Negri I (2002) Non-linear statistical modeling of high frequency ground ozone data. Environmetrics 13:225–241

Fiester U, Balzer K (1991) Surface ozone and meteorological predictors on a subregional scale. Atmos Environ 25:1781–1790

Hansen BE (1999) Threshold effects in non-dynamic panels: estimation, testing, and inference. J Econom 93:345–368

Huang L-S, Smith R (1999) Meteorologically-dependent trends in urban ozone. Environmetrics 10:103–118

Kim SE, Kumar A (2005) Accounting seasonal nonstationarity in time series models for short-term ozone level forecast. Stoch Env Res Risk A 19:241–248

Kumar A, Bellam N, Sud A (1999) Performance of industrial source complex model in predicting long-term concentrations in an urban area. Environ Prog 18(2):93–100

Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models. Biometrika 65:297–303

Millionis M, Davies TD (1994) Regression and stochastic models for air pollution-I. Review, comments and suggestion. Atmos Environ 28(17):2801–2810

Monti AC (1994) A proposal for residual autocorrelation test in linear models. Biometrika 81:776–780

Potter SM (1995) A nonlinear approach to US GNP. J Appl Econom 10:109–125

Robeson SM, Steyn DG (1990) Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. Atmos Environ 24B:303–312

Schwarz F (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Thomson ML, Reynolds J, Lawrence HC, Guttorp P, Sampson PD (2001) A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmos Environ 35:617–630

Tiao GC, Tsay RS (1994) Some advances in non-linear and adaptive modelling in time series. J Forecast 13:109–131

Tong H (1983) Threshold models in non-linear time series analysis. In: Brillinger D, Fienberg J, Gani J, Hartigan J, Krickberg K (eds) Lecture notes in statistics. Springer, Heidelberg

Tong H (1990) Nonlinear time series: a dynamic system approach. Oxford University Press, Oxford

Venables WN, Ripley BD (2002) Modern applied statistics with S. Fourth Ed. Springer, Heidelberg