文章编号:1004-4574(2007)03-0096-04

非线性时序法在城市大气污染预测中的应用

张建同,尤建新

(同济大学 经济与管理学院,上海 200092)

摘要:建立城市大气污染预测模型是治理城市大气污染的重要工作。在简述时间序列方法基本原理的基础上,分析了系数为变量的自回归滑动平均(ARMA)模型、截断ARMA模型,和残差为自回归综合滑动平均(ARIMA)的半参数方法等城市大气污染预测模型。以法国某城市为例,分别采用AR模型和系数为变量的AR模型对大气污染进行了预测。通过比较预测结果可知,基于非线性时间序列方法的城市大气污染预测模型可以提高预测精度,降低预测误差。

关键词:大气污染;预测;非线性时间序列;半参数估计

中图分类号:X51 **文献标识码**:A

Application of nonlinear time sequence method to prediction of atmospheric pollution in City

ZHANG Jian-tong, YOU Jian-xin

(School of Economics and Management, Tongji University, Shanghai 200092, China)

Abstract: It is important to propose reasonable models for foretelling the atmosphere pollution. After reviewing the fundamental theory of time serial methods, variable parameter autoregressive moving average(ARMA) model, truncation ARMA model and remnant ARIMA semi-parameter model were analyzed. A comparison was made between experimental and theoretical results obtained by autoregressive(AR) and variable parameter AR model respectively. It seems that the nonlinear time serial model can be used to forecast the atmosphere pollution.

Keywords: atmospheric pollution; prediction; nonlinear time sequence; semi-parameter estimation

城市大气污染问题是一个严重的环境污染问题。"伦敦烟雾事件"[1]、"光化学烟雾事件"[2]、酸雨和沙尘暴[2]等一系列事件显示,城市大气污染严重影响社会正常生产和生活,甚至危机人类生存。控制大气污染、提高空气质量日益成为世界各国环境综合治理的目标。作为城市大气污染治理的重要工作之一,城市大气污染预测[3]有利于建立城市大气污染预报体系,有效降低大气污染治理成本。建立合理精确的预测模型是城市大气污染预测的基础。

大气污染预测的方法模型主要有数值预测方法和统计预测方法。数值预测方法是掌握大气污染物在空气中演变规律,以天气形势及其气象要素指标为依据,定量描述空气中大气污染物的浓度,对未来大气环境质量状况进行定性或定量的分析[3-4]。统计预测方法是不依赖物理、化学及生物过程,通过分析事件规律来进行预测的方法[5-8]。作为统计预测方法之一,时间序列方法已经被用于大气污染预测,但大部分仅使用线性时间序列模型,非线性时间序列方法预测城市大气污染仍比较少[6-7,9]。

收稿日期:2006-12-10; **修订日期**:2007-05-10

基金项目:国家自然科学基金资助项目(70640007)

作者简介:张建同(1966-),女,副教授,主要从事应用统计研究. E-mail:zhangjiant@163.com

1 时间序列方法的基本原理

时间序列分析依据随机过程理论和数理统计学,研究随机数据序列所遵从的统计规律,以解决实际问题。其包括一般统计分析,统计模型的建立与推断,以及关于随机序列的最优预测、控制和滤波等。

简单的时间序列模型,即自回归模型 AR(1)的形式为

$$X_t = aX_{t-1} + \varepsilon_t, \quad t \in Z \tag{1}$$

式中,a 是实数;ε_t 是白噪声,即一列不相关随机序列,均值为零,方差为 σ^2。当 $EX_t = \mu$ 时,对任意的 $t \in Z$,式(1)可写成

$$X_t - \mu = a(X_{t-1} - \mu) + \varepsilon_t, \quad t \in Z \tag{2}$$

式中,a 是正数,其估计为 $\hat{a}$

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_-)(x_{i+1} - \bar{x}_+)}{[\sum_{i=1}^{n}(x_i - \bar{x}_-)^2 \sum_{i=2}^{n}(x_{i+1} - \bar{x}_+)^2]^{1/2}} \tag{3}$$

式中,$x_1, \cdots, x_n$ 是观测值,$\bar{x}_-$ 是 $x_1, \ldots, x_{n-1}$ 的平均值,$\bar{x}_+$ 是 $x_2, \ldots, x_n$ 的平均值。

自回归模型 AR(1)的一般形式,即自回归滑动平均(autoregressive moving average)ARMA(p,q)的形式为

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t + \sum_{j=1}^{q} b_j \varepsilon_{t-j} \tag{4}$$

式中,a_i 和 b_j 取值为区间(-1,1)的数,ε^t 是白噪声。如果下列条件成立,则可认为时间序列 ARMA(p,q)平稳。

i) $E|X_t|^2 < \infty$, $t \in Z$

ii) $EX_t = m$, $t \in Z$

iii) $\mathrm{Cov}(X_r, X_s) = \mathrm{Cov}(X_{r+t}, X_{s+t})$, $r,s,t \in Z$

对于平稳的时间序列,对于所有的 $t,h \in Z$,自协方差函数可表示为 $r_X(h) = \mathrm{Cov}(X_{t+h}, X_t)$,自相关函数可表示为 $\rho_X(h) = r_X(h)/r_X(0) = \mathrm{Corr}(X_{t+h}, X_t)$

从现象的所得的原始数据往往是不平稳的。例如,几年内每天的最高温度为一个时间序列 X_t。由于大气温室效应和季节交替,气温呈趋势性的逐渐上升和季节性的周期变化,X_t 的数学期望不可能为常数。

在将原始数据转换为平稳序列之前,需首先对原始数据作变换,目的是使方差稳定。Graf - Jacott (1993)[11]对于正数数据采用了 Box - Cox 转换

$$U_t = \begin{cases} \dfrac{X_t^{\lambda} - 1}{\lambda}, & X_t \geq 0, \ \lambda > 0 \\ \ln X_t, & X_t > 0, \ \lambda = 0 \end{cases} \tag{5}$$

式中 λ 是一个参数。

将非平稳序列变换成为平稳序列,通常采用两种方法:

i)将非平稳时间序列表示为

$$X_t = m_t + s_t + Y_t \tag{6}$$

式中,m_t 是趋势变量,s_t 是季节性变量,Y_t 是平稳的时间序列。若只有趋势变量,可采用最小二乘的方法估计 m_t;

ii)采用 Box - Jenkins(1976)[10]的方法,对以下时间序列

$$Y_{t,1} = (X_t - X_{t-1})$$

$$Y_{t,2} = (Y_{t,1} - Y_{t-1,1})$$

依次运算 d 步至获得平稳的 ARMA(p,q)模型,该模型称为 ARIMA(p,q,d)(Autoregressive Integrated Moving Average)。此外,当时间序列呈季节性变化时,可以采用 SARIMA 模型(seasonal ARIMA),其模型表达式比较复杂,详细论述参见文献[7]。

用 AR(p)模型进行预测的方式很简单,只需要将噪声项去掉,在时间点 $t+1$ 的预测值是以往数值的线性组合,但通常需给出预测值的置信区间。

2 城市大气污染预测的模型

由于造成污染的原因复杂,线性模型无法很好地评价污染的程度,学者们提出了系数为变量的 ARMA

模型。Barrat 等(1990)[6]用该模型对空气中的 NO 进行了预测;近几年,一些非参数模型和半参数方法也被用于大气污染的预测[7,9]。

2.1 系数为变量的 ARMA 模型

为了避免数据的非平稳性,设时间序列 $X_h(J)$ 为每一天同一时刻 h 的数据,系数 $a_{i,h}$ 随着时间变化的 AR(p)模型为

$$X_h(J) = \sum_{i=1}^{p} a_{i,h} X_{h-i}(J) + e_h(J) \tag{7}$$

此模型可反映数据的昼夜变化和季节变化。采用最小二乘法估计系数 $a_{i,h}$,可得到 f 小时以后的预测值 $X_{h+f}(J)$。

2.2 截断 ARMA 模型

模型 AR(1)的截断形式为

$$X_t = \begin{cases} a_1' X_{t-1} + \varepsilon_t & \text{如果 } X_t < \alpha \\ a_2^* X_{t-1} + \varepsilon_t & \text{如果 } X_t \geq \alpha \end{cases} \tag{8}$$

式中,α 是阈值,a_1' 和 a_2^* 是系数,ε_t 为噪声。

上述模型可以被推广为一般的形式,记作 TAR($l, p_1, \ldots, p_l$)(Threshold Auto Regressive)。此时需要引入另一个时间序列(Y_t),该序列可以决定(X_t)的特征,即对于所有的 t,当 $\alpha_{i-1} \leq Y_t < \alpha_i, i = 1, \ldots l$

$$X_t = a_{0i} + \sum_{j=1}^{p_i} a_i(j) X_{t-j} + \varepsilon_t \tag{9}$$

式中,$\alpha_0 = -\infty < \alpha_1 < \ldots < \alpha_{t-1} < \alpha_t = \infty$ 是阈值。

Melard 和 Roy(1998)[9]采用了更一般的模型,对于 $\alpha_{i-1} \leq Y_t < \alpha_i, i = 1, \ldots l$

$$X_t = \sum_{j=1}^{p_i} a_i(j)(X_{t-j} - \mu_{I(t-j)}) + \varepsilon_t - \sum_{k=1}^{q} b_i(k)\varepsilon_{t-k} \tag{10}$$

式中,当 $\alpha_{i-1} \leq Y_t < \alpha_i$ 时,$I(t) = i$。截断 ARMA 模型的详细介绍可参见 Tong(1983)[9]。

2.3 残差为 ARIMA 的半参数方法

假定预测时间为半小时,每隔 5 分钟提取一次数据,预测第 $t+6$ 时刻的值。为了避免估计太多参数,可采用半参数的方法。令(X_t)为一个随机过程,可以用 Nadaraya - Waston 回归的方法估计 $E(X_{t+6}/X_t, X_{t-1})$,记为 $\hat{E}(X_{t+6}/X_t, X_{t-1})$。

由于残差 $\hat{Z}_i = X_i - \hat{E}(X_i/X_{i-6}, X_{i-7})$ 不一定是一个白噪声,通常可用 ARIMA 对 $\hat{Z}_{t-6}, \ldots, \hat{Z}_t$ 建模,得到预测值 $\hat{Z}_{t+6}$。$\hat{X}_{t+6}$的预测值为 $\hat{E}(X_{t+6}/X_t, X_{t-1}) + \hat{Z}_{t+6}$。半参数方法的预测结果好于非参数方法[7,9],且远好于 ARIMA 方法[7,9]。

3 模拟结果及讨论

以法国某城市 1995 年 7 月、8 月和 9 月的空气中臭氧的浓度((g/m^3)为观测数据,数据采集间隔为 1 h,预测长度为 2 h,分别采用古典 AR 模型和模型(7)对大气中的 NO 进行了预测。

用 Yule - Walker 方程求 $a_1, \ldots, a_p$ 的估计 $\hat{a}_1, \ldots, \hat{a}_p$

$$\begin{aligned} r_1 &= \hat{a}_1 + \hat{a}_2 r_1 + \hat{a}_3 r_2 + \ldots + \hat{a}_p r_{p-1} \\ r_2 &= \hat{a}_1 r_1 + \hat{a}_2 + \hat{a}_3 r_1 + \ldots + \hat{a}_p r_{p-2} \\ r_3 &= \hat{a}_1 r_2 + \hat{a}_2 r_1 + \hat{a}_3 + \ldots + \hat{a}_p r_{p-3} \\ r_p &= \hat{a}_1 r_{p-1} + \hat{a}_2 r_{p-2} + \hat{a}_3 r_{p-3} + \ldots + \hat{a}_p \end{aligned} \tag{11}$$

式中 r_p 是经验自相关系数。可以得到白噪声方差的估计 $S_a^2(p)$。

为了避免 p 过大,采用下面两个准则,即选择使下边两式达到最小的 p

$$B_{IC}(p) = n\ln\left[\frac{n}{n-p-1} s_a^2(p) + (p+1)\ln n\right] \tag{12}$$

$$A_{IC}(p) = n\ln\left[\frac{n}{n-p-1} s_a^2(p) + 2(p+1)\right] \tag{13}$$

式中 n 为样本容量的大小。

当 AR(p)模型中的系数为变量时,Yule - Walker 方程为:

$$\begin{aligned} r_{1,h} &= \hat{a}_{1,h} + \hat{a}_{2,h}r_{1,h} + \hat{a}_{3,h}r_{2,h} + \cdots + \hat{a}_{p,h}r_{p-1,h} \\ r_{2,h} &= \hat{a}_{1,h}r_{1,h} + \hat{a}_{2,h} + \hat{a}_{3,h}r_{1,h} + \cdots + \hat{a}_{p,h}r_{p-2} \\ r_{3,h} &= \hat{a}_{1,h}r_{2,h} + \hat{a}_{2,h}r_{1,h} + \hat{a}_{3,h} + \cdots + \hat{a}_{p,h}r_{p-3} \\ r_{p,h} &= \hat{a}_{1,h}r_{p-1,h} + \hat{a}_{2,h}r_{p-2} + \hat{a}_{3,h}r_{p-3} + \cdots + \hat{a}_{p,h} \end{aligned} \tag{14}$$

对式(14)求解,可得到模型(7)中的参数 $a_{1,h}, a_{2,h}, \cdots, a_{p,h}$ 的估计值 $\hat{a}_{1,h}, \hat{a}_{2,h}, \cdots, \hat{a}_{p,h}$。通过 $A_{IC}(p)$(12)和 $B_{IC}(p)$(14)的计算,选择 $p=1$ 最合适。两种结果的预测结果见图 1 和图 2,两种方法的残差诊断结果见表 1。直观上可以看出模型(7)更加符合实际情况,尤其是它对出现尖峰的情况,即对污染度突然上升时的预测能力比较强。通过表 1 的两种情况的残差对比及两种模型的标准差的对比,可知系数为变量的 AR 模型比传统的 AR 模型预测精度高。

表 1　残差诊断对比表

Table 1　Contrast of residual diagnosis

	AR(1)	系数为变量的 AR(1)
残差数	329	329
调正残差平方和	10 026.576	9 323.227
残差平方和	10 042.133	9 860.321
残差方差	30.756	28.600
模型标准误差	5.546	5.348

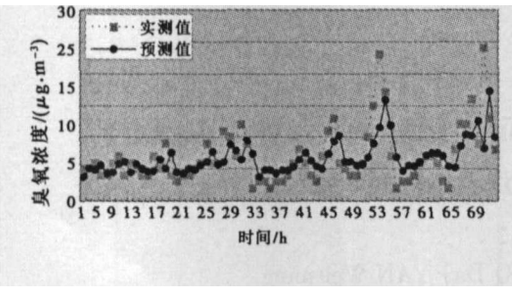

图 1　用 AR(p)模型对空气中臭氧的预测结果

Fig. 1　Predicted result of ozonic density in air using AR(p) model

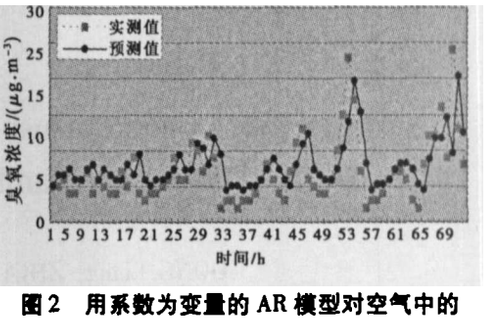

图 2　用系数为变量的 AR 模型对空气中的臭氧浓度预测结果

Fig. 2　Predicted result of ozonic density in air using AR model with coefficients as variable

4　结语

传统的 AR、ARMA 模型可近似预测城市大气污染,但缺乏足够的预测精度。系数为变量的时间序列模型对城市大气污染的预测结果与实际数据吻合较好,其方法优于线性时间序列模型。非线性时间序列方法应用于城市大气污染,能够提高预测精度,降低预测误差,为污染治理提供决策依据,具有广泛的适用性。

参考文献:

[1]　李浩,奚旦立,唐振华,等. 英国大气污染控制及行动措施[J]. 干旱环境监测,2005,19(1):29 - 32.

[2]　李卫民. 城市的大气污染与防治[J]. 城市之光,2000,5:87 - 88.

[3]　刘振忠,董芃,王丽. 城市大气环境污染预测与容量控制方法研究[J]. 电站系统工程, 2005, 7:17 - 19.

[4]　吴晓鸣. 城市大气污染预测简介[J]. GANSU METEOROLOGY,2000,1:24 - 54.

[5]　刘永,郭怀成. 城市大气污染物浓度预测方法研究[J]. 安全与环境学报,2004,8:60 - 62.

[6]　Barrat M, Lecluse Y, Slamani Y. Etude comparative de différents, modèles, mathématiques pour la prédiction des niveaux de pollution atmosphèrique, analyse univariable[J]. RAIRO APII, 1990,24(3):283 - 298.

[7]　Gonzalez - Manteiga W, Prada - Sanchez J M, Cao R, et al. Times series analysis for ambient concentrations, Atmospheric[J]. Environment, 1993,27A(3):153 - 158.

[8]　Guillas S, Rhomari N, Zhang J. Bilan de l'existant en matiere de prevision statistique des pics de pollution[R]. Paris: ISUP,2000:15 - 25.

[9]　Prada - Sanchez J M, Febrero - Bande M, Coto - Yanez T, et al. Prediction of SO_2 pollution incident near a power station using partially linear models and an historical matrix of predictor - reponse vectors[J]. Environmetrics,2000,11:209 - 225.

[10]　Box G E, Jenkins G M. Times series analysis. Forecasting and control[M]. San Francisco: Holden Day, 1976:123 - 146.

[11]　M. Graf - Jacottet. A flexible model for ground ozone concentration[J]. Environmetrics 4, 1993:23 - 37.