

一类非线性模型建模方法及其在铁路客流量预报上的应用

钟秋海 黎新华 张志方

(北京工业学院)

摘 要

本文概述了时序分析的非线性模型类,详细叙述了门槛自回归模型的建模思想和方法。对于铁路客流量数据进行了建模和预报,逐月预报1982年和1983年的准确度分别在91.9%和91.4%以上。这对于加强中国铁路管理的科学性有很大帮助。这个方法同样可以推广应用于其他具有周期性数据的业务部门。

一、前 言

自从Box和Jenkins总结出一套时序建模方法以后,时序分析法得到了广泛的应用。吴贤铭又从系统角度出发提出了动态数据系统方法,它有一维寻优、计算量少,可根据定阶准则由计算机全自动建模的优点,因而工程上使用方便。

但是,这些方法都是用线性模型,而实际的系统常常是非线性的。用线性模型去逼近常不能得到令人满意的结果,因而有人不断提出非线性模型用于时间序列分析。其中有汤家豪博士提出的门槛自回归模型^[1] (Threshold Autoregressive Model),其一般形式为

$$X_t = \varphi_0^{(j)} + \sum_{i=1}^{m_j} \varphi_i^{(j)} X_{t-i} + a_t^{(j)}, \text{ 当 } \gamma_{j-1} < X_{t-d} \leq \gamma_j$$
$$j=1, 2, \dots, J, \quad (1)$$

其中 $-\infty < \gamma_0 < \gamma_1 < \gamma_2 < \dots < \gamma_J < \infty$ 为门槛值, d 为延迟参数。当 X_{t-d} 落在 γ_{j-1} 与 γ_j 之间时,就用第 j 个自回归模型(阶数为 m_j)。 $\{a_t^{(j)}\}$ 对每个固定的 j 是方差为 σ_j^2 的白噪声,且 $j' \neq j''$ 时 $\{a_t^{(j')}\}$ 与 $\{a_t^{(j'')}\}$ 是相互独立的。模型用 SETAR($J; m_1, m_2, \dots, m_J$) 表示。它可以有效地描述存在极限环或极限点的非线性系统,并且可以刻画幅频依

赖和共振跳跃现象。最常用到的模型 $J=2$ 。

Ozaki, T. 提出了指数自回归模型^[2] (Exponential Autoregressive Model) 其一般形式为

$$X_t = \sum_{i=1}^p (\varphi_i + \pi_i e^{-\gamma X_{t-1}^2}) X_{t-i} + a_t, \quad (2)$$

其中 φ_i, π_i 都是常数, $\gamma > 0$ 。它也可以刻画幅频依赖和共振跳跃现象。

Jones, D. A. 研究了更广泛的一类非线性自回归模型^[3] (Nonlinear Autoregressive Model), 一般形式为

$$X_t = \lambda(X_{t-1}, \dots, X_{t-p}) + a_t, \quad (3)$$

函数 λ 的形式依情况而异。门槛自回归和指数自回归可以作为它的两个特例。

Granger, C. W. J. 和 Andersen, A. P. 最早提出了双线性模型^[4] (Bilinear Model), 其一般形式为

$$X_t = \sum_{j=1}^p \varphi_j X_{t-j} + \sum_{i=0}^q \theta_i a_{t-i} + \sum_{k=0}^Q \sum_{l=1}^P \beta_{kl} a_{t-k} X_{t-l}. \quad (4)$$

由于还没有识别双线性模型的强有力的判别准则, 它的应用目前只限于最简单的几种。

此外, 还有 Priestly, M. B. 着重研究的状态依赖模型^[5] (State-Dependent Model), 表达式为

$$X_t = \mu(X_{t-1}) + \sum_{i=1}^p \varphi_i(X_{t-1}) X_{t-i} + \sum_{j=0}^q \psi_j(X_{t-1}) \epsilon_{t-j}, \quad (5)$$

其中 $X_{t-1} = (X_{t-1}, X_{t-2}, \dots, X_{t-h})^T$ 。参数都是状态 X_{t-1} 的函数。其模型识别就更加困难。

在这些非线性模型类中, 工程应用比较成功的是门槛自回归模型。

二、门槛自回归模型的建模和预报

(1) 一个时间序列可以看作一个系统的输出。当这个序列是周期性数据时, 则可以认为它是一个非线性系统的极限环。非线性系统最简单的极限环在相平面上只有一条切换线。在这切换线的一侧用一个线性模型描述, 在另一侧则用另一个线性模型描述。类似地, 在时序分析中采用门槛的概念。采样值大于某个值 (称门槛) 时采用一个线性模型 $AR(P_1)$ 描述, 小于门槛时则采用另一个线性模型 $AR(P_2)$ 描述。系统中常见的非线性特性是死区、间隙、时滞、饱和。在时序分析中采用延迟的概念。即根据延迟 d 步后的 X_{t-d} 是大于还是小于门槛决定采用哪个模型来描述。这就是门槛自回归模型拟合非线性系统的基本思想。由于门槛自回归能很好地拟合周期性数据, 刻划出非线性系统的极限

或极低点,建模容易,因而推广应用起来方便。

门槛自回归模型的建模步骤是(以分段数 $I=2$ 为例):

1. 对原始数据进行变换(如取对数),去掉均值或趋势项。

2. 给定延迟步数 d 和每段中 AR 模型最大可能的阶数 L , 令 $n_0 = \max(d, L)$ 。

3. 设 $TQ = \{t_{q1}, t_{q2}, \dots, t_{qi}\}$ 为门槛备择集合。设序列 $\{x_1, x_2, \dots, x_N\}$ 按 $x_{t-d} (t > n_0)$ 大于还是小于 t_{q1} 分为两个序列 $\{x_{n_0+1}, x_{n_0+4}, x_{n_0+5}, \dots\}$, $\{x_{n_0+2}, x_{n_0+3}, x_{n_0+6}, \dots\}$, 则分段模型为

$$\begin{pmatrix} x_{n_0+1} \\ x_{n_0+4} \\ x_{n_0+5} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & x_{n_0} & x_{n_0-1} & \dots & x_{n_0-k_1} \\ 1 & x_{n_0+3} & x_{n_0+2} & \dots & x_{n_0+4-k_1} \\ 1 & x_{n_0+4} & x_{n_0+3} & \dots & x_{n_0+5-k_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \varphi_0^{(1)} \\ \varphi_1^{(1)} \\ \varphi_2^{(1)} \\ \vdots \\ \varphi_{k_1}^{(1)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{n_0+1}^{(1)} \\ \varepsilon_{n_0+4}^{(1)} \\ \varepsilon_{n_0+5}^{(1)} \\ \vdots \end{pmatrix}$$

$$\text{即} \quad X_1 = A_1 \Theta_1 + \varepsilon_1, \quad (6)$$

$$\begin{pmatrix} x_{n_0+2} \\ x_{n_0+3} \\ x_{n_0+6} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & x_{n_0+1} & x_{n_0} & \dots & x_{n_0+2-k_2} \\ 1 & x_{n_0+2} & x_{n_0+1} & \dots & x_{n_0+3-k_2} \\ 1 & x_{n_0+5} & x_{n_0+4} & \dots & x_{n_0+6-k_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \varphi_0^{(2)} \\ \varphi_1^{(2)} \\ \varphi_2^{(2)} \\ \vdots \\ \varphi_{k_2}^{(2)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{n_0+2}^{(2)} \\ \varepsilon_{n_0+3}^{(2)} \\ \varepsilon_{n_0+6}^{(2)} \\ \vdots \end{pmatrix}$$

$$\text{即} \quad X_2 = A_2 \Theta_2 + \varepsilon_2. \quad (7)$$

用线性最小二乘法估计出各段模型的参数。

4. 在门槛取值为 t_{q1} 时, 求出使各段模型 AIC 最小的阶数 $\hat{K}_1^{(1)}$, $\hat{K}_2^{(1)}$ 。即

$$\text{AIC}(\hat{K}_j^{(1)}) = \min_{0 < K_j \leq L} \{N_j \ln [Rss_j(K_j)/N_j] + 2(K_j + 1)\}$$

$$j = 1, 2, \quad (8)$$

其中 $Rss_j(K_j)$ 为第 j 段 K_j 阶自回归模型的残差平方和。 N_j 为其子样本量。并求得总的 AIC:

$$\text{AIC}(t_{q1}) = \text{AIC}(\hat{K}_1^{(1)}) + \text{AIC}(\hat{K}_2^{(1)}). \quad (9)$$

5. 门槛取其他值, 重复 3, 4 过程。由此选择出延迟为 d 时的最佳门槛 $\hat{t}_q$ 。即

$$AIC(\hat{t}_q) = \min_{t_{qi} \in TQ} \{AIC(t_{qi})\} \quad (10)$$

6. 改变延迟参数 d 的取值, 重复 2~5 的过程, 由 AIC 极小定出最佳延迟 $\hat{d}$. 到此, 定出了最佳延迟步数 $\hat{d}$, 最佳门槛值 $\hat{t}_q$ 及其对应的各段模型最佳阶数 $\hat{K}_1$, $\hat{K}_2$ 和参数 θ_1 , θ_2 . 建模过程结束.

这里需说明几点. 在第 3 步中, Tong, H. 建议采用的门槛备择集合为 $(x_{\max} - x_{\min})$ 的 30%, 40%, 50%, 60% 和 70%, 这样做比较粗糙. 经大量计算表明, $AIC(t_q)$ 是 t_q 的单峰函数^[6]. 因此, 我们采用黄金分割法来优选门槛值. 用两分点之间距离小于某一给定数值来控制停止继续分割, 得到了更好的效果.

对于门槛不同, 使得大于和小于门槛的子样本数有可能相差悬殊, 因此当某子样本少时, 应适当降低最高阶数的限制, 这样可以减少计算量.

提前 i 步预报 x_{N+i} 时, 要判别 x_{N+i-d} 是属于哪个范围, 从而决定用哪个模型来预报 x_{N+i} . 所以, 当预报步数 $i \leq d$ 时, x_{N+i-d} 都是原始数据, 因而判别是大于还是小于门槛值时比较客观, 预报也比较准确. 当 $i > d$ 时, 则要根据预报数据进行判别采用哪个模型外推了, 相对误差要大一些. 这样做出的长期预报函数若是周期的, 则说明系统存在稳定的极限环; 若趋于一常数, 则说明系统存在极限点.

三、应用于铁路客运量预报

目前铁道部及其下属机构只能根据以往的统计数据凭经验作出近期(半月至一个月)的预计, 或在三季度末作出下一年全年运输量的预计, 以作为下一年度运输计划的依据. 工作人员不可能对下一年度每个月的运输量作出预计.

现在我们对统计数据进行分析. 给出 1971 年至 1981 年的逐月客运数据图如图 1 所示. 由图可以看出, 全国铁路运输量随着国民经济的发展逐年增加, 而且一年 12 个月当中呈现出一个高峰二个低谷的变化周期.

我们在 1982 年一季度利用 1971 年 1 月至 1981 年 12 月的 132 个数据进行建模和预报. 建模结果, 确定性趋势项为

$$(1 + (K)S + (M)S^2) \cdot 6.13391 + 0.00423t \quad t = 1, 2, 3, \dots \quad (11)$$

去掉趋势项以后建模结果为 SETAR(2; 14, 1), 延迟 $\hat{d} = 12$, 门槛值 $\hat{t}_q = 0.083667$. 模型参数如表 1 所示.

利用这个模型对 1982 年和 1983 年直至 1986 年的客运量作了长期预报, 经过二年的考核, 预报值与实际数据非常接近. 最大的月预报相对误差 1982 年为 8.1%, 1983 年为 8.6%, 所以准确率分别为 91.9% 和 91.4% 以上. 详细情况见表 2 和图 2.

表 1

小于门槛段 (14 阶)			大于门槛段 (1 阶)		
	参 数 值	95% 置信限		参 数 值	95% 置信限
φ_0	0.00757	0.0054	φ_0	0.03386	0.0284
φ_1	0.54847	0.1012	φ_1	0.16648	0.3385
φ_2	0.02288	0.1025			
φ_3	0.0346	0.0932			
φ_4	-0.13552	0.0910			
φ_5	0.04493	0.0917			
φ_6	-0.06281	0.0882			
φ_7	-0.08222	0.0926			
φ_8	-0.15202	0.0882			
φ_9	0.07209	0.0893			
φ_{10}	-0.11825	0.0913			
φ_{11}	0.06824	0.0981			
φ_{12}	0.65885	0.1255			
φ_{13}	-0.28082	0.1014			
φ_{14}	-0.27546	0.0959			
$\text{Var } \varepsilon_i^{(1)} = 0.001397$			$\text{Var } \varepsilon_i^{(2)} = 0.004689$		
平均 $\text{Var } \varepsilon_i = 0.001759$					

表 2

年、月	模型 预报值	真实值 (万人)	相对误差 (%)	准确率 (%)
82.1	8304.39	9037	-8.11	91.9
82.2	8365.30	8058	3.81	96.2
82.3	8265.47	8329	-0.76	99.2
82.4	8361.44	8220	1.72	98.3
82.5	8406.27	8183	2.73	97.3
82.6	7814.27	7348	6.34	93.7
82.7	8329.24	8120	2.58	97.4
82.8	8885.20	8844	0.47	99.5
82.9	8333.33	7995	3.86	96.1
82.10	8465.93	8118	4.29	95.7
82.11	8336.78	8076	2.86	97.1
82.12	8089.36	8560	-5.5	94.5

续表 2

全 年	99732.14	98888	0.85	99.1
83.1	8730.06	8902	-1.93	98.1
83.2	8850.85	8978	-1.42	98.6
83.3	8741.69	9134	-4.30	95.7
83.4	8914.76	8864	0.57	99.4
83.5	8944.61	8916	0.32	99.7
83.6	8487.32	7875	7.78	92.2
83.7	8944.99	8612	3.87	96.1
83.8	8304.46	9445.5	-1.49	98.5
83.9	8850.63	8411	5.23	94.8
83.10	9038.39	8321	8.62	91.4
83.11	8825.41	8500	3.83	96.2
83.12	8606.49	8988	-4.24	95.8
全 年	106239.66	104946	1.23	98.8

续表 2

年、月	模型预报值	真 实 值	年、月	模型预报值	真 实 值
84.1	9125.45		85.8	10242.20	
84.2	9246.34		85.9	10015.12	
84.3	9159.11		85.10	10209.72	
84.4	9386.25		85.11	9954.78	
84.5	9417.69		85.12	9740.87	
84.6	9090.91		全 年	118283.88	
84.7	9523.61		86.1	10039.02	
84.8	9759.47		86.2	10058.61	
84.9	9419.76		86.3	9986.30	
84.10	9609.60		86.4	10226.43	
84.11	9368.89		86.5	10252.29	
84.12	9145.08		86.6	10142.21	
全 年	112252.16		86.7	10566.65	
85.1	9549.24		86.8	10724.99	
85.2	9631.28		86.9	10615.29	
85.3	9561.48		86.10	10832.86	
85.4	9812.86		86.11	10614.00	
85.5	9847.78		86.12	10407.56	
85.6	9640.99		全 年	124466.21	
85.7	10067.56				

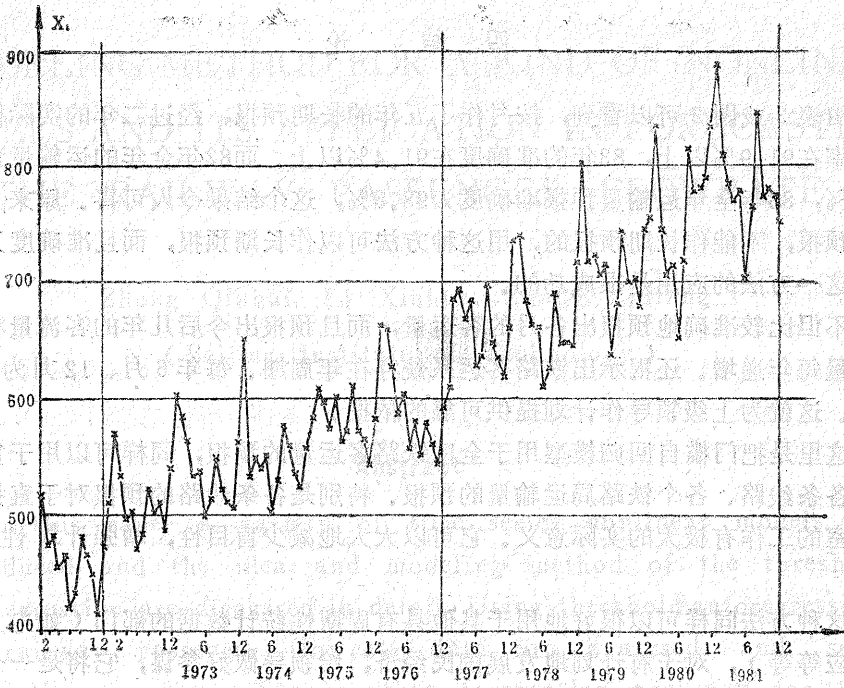


图 1 全国铁路客流量统计曲线

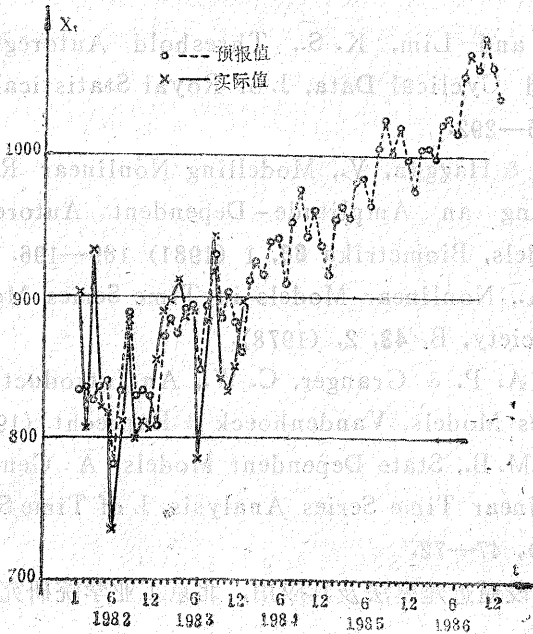


图 2 预报曲线

四、结 论

1. 由表 2 或图 2 可以看到, 按月作了五年的长期预报, 经过二年的实际检验, 82 年的准确率在 91.9% 以上, 83 年的准确度在 91.4% 以上. 而 82 年全年的运输量预报准确度为 99.1%, 83 年全年运输量预报准确度为 98.8%, 这个结果令人可喜. 原来只能凭经验作近期预报, 不能作长期预报的, 用这种方法可以作长期预报, 而且准确度又是那么高, 说明这一方法的应用是很成功的.

2. 不但比较准确地预报出各月的客运量, 而且预报出今后几年的客流量将按六千万人的数量每年递增. 还揭示出铁路客运系统存在年周期, 每年 6 月、12 月为低潮, 8 月为高峰. 这就为上级领导作计划提供可靠的依据.

3. 这里是把门槛自回归模型用于全国铁路客运量的预报, 同样可以用于货运量的预报以及各条线路、各个铁路局运输量的预报. 特别是各条线路的预报对于直接指导铁路局调度室的工作有极大的实际意义. 它可以大大地减少盲目性, 增强计划性和科学性.

4. 这种方法同样可以很好地用于其他具有周期性统计数据的部门 (如电力部门、自来水供应等等), 对于有计划地发展国民经济, 给领导做好参谋, 它将是一个有力的工具.

参 考 文 献

- [1] Tong, H. and Lim, K. S., Threshold Autoregression, Limit Cycles and Cyclical Data, J. of Royal Statistical Society, B, **42** (1980), 245—292.
- [2] Ozaki, T. & Haggan, V., Modelling Nonlinear Random Vibrations Using an Amplitude-Dependent Autoregressive Time Series Models, Biometrika **68**, 1 (1981) 189—196.
- [3] Jenes, D. A., Nonlinear Models in Time Series Models, J. R. Statistical Society, B, **43**, 2, (1978).
- [4] Andersen, A. P. & Granger, C. W., An Introduction to Bilinear Time Series Models, Vandenhoeck & Ruprecht (1978).
- [5] Priestley, M. B., State Dependent Models: A General Approach to Non-linear Time Series Analysis, J. of Time Series Analysis, **1**, 1 (1980), 47—72.
- [6] 黎新华, 逐段逼近建模法及其应用, 北京工业学院研究生论文 (1982).

A MODELING METHOD FOR A KIND OF NONLINEAR MODELS AND ITS APPLICATION IN FORECASTING THE RAILWAY PASSENGER FLOWRATE

Zhong Qiuhai, Li Xinhua, Zhang Zhifang

(Beijing Institute of Technology)

Abstract

In this paper a variety of time series nonlinear models are introduced and the idea and modeling method of the threshold autoregression are discussed in detail. Using threshold autoregression, the railway passenger flowrate data are analysed and some forecasting are given. The monthly forecasting accuracy for 1982 and 1983 is greater than 91.9% and 91.4% respectively. This will be of great help in running China's railway in a more scientific way. The method can also be applied to those fields in which periodic data often exist.