ARMA模型和TAR模型在貂年捕获量中的预测结果比较

胡一睿[1]，曲荣华[1]，胡潇[1]，王雪[2]
（1. 北京师范大学数学科学学院，北京 100875;
2. 北京师范大学物理系，北京 100875）

摘要：[目的] 对 ARMA 模型和门限自回归模型在貂年捕获量的预测结果进行比较。[方法] 利用我国 1848 年至 1909 年貂年捕获量数据，分别建立 ARMA 模型和门限自回归模型，拟合后对未来五年的年捕获量进行预测，比较 2 个模型的拟合和预测效果。[结果] ARMA 模型和门限自回归模型的平均误差率(MER)分别为 27.93％、3.2％。[结论] 门限自回归模型对于处理循环的数据具有明显的预测优势，预测效果优于 ARMA 模型，对解决时间序列类型的资料有很好的实用价值。

关键词： 时间序列；门限自回归；ARMA 模型；平均误差率
中图分类号：O21

The comparison of ARMA and TAR model in the prediction of time series data

HU Yirui, QU Ronghua, HU Xiao, WANG Xue
(School of Mathematical Sciences Beijing Normal University, Beijing 100875)

Abstract: [Objective] Compare ARMA model and TAR model in the predicted results of data. [Method]With data from 1848 to 1909, we established ARMA model TAR model separately, aand then compared two model fitting and predicting effects. [Results]MER in ARMA models and TAR model is 27.93%, 3.2%. [Conclusion]TAR model has obvious advantages in cycling data , when compared with ARMA model in solving time series data and has a good practical value.

Key words: Time series; TAR; ARMA; MER

0 引言

ARMA 模型是研究时间序列的重要方法，由自回归模型（简称 AR 模型）与滑动平均模型（简称 MA 模型）为基础“混合”构成。TAR 模型用于处理具有循环性质的加拿大山猫数据，取得了较好的效果[1]。通过观察时间序列数据 ，发现其序列恰是一种循环，由于循环的循环节长度不固定，因此不能将其看作周期模型。在循环中，捕获量的增加期和减少期是交替出现的，其类似于生态学中的捕食关系。当捕获量经过一段增加时期后，由于貂的总体数量较少，导致会有一段时期难以捕获，因此，捕获量也就会减少。而捕获量减少后，貂的数量有一个累积的过程，当貂的数量恢复到一定程度后，捕获量又开始增大，如此过程，循环交替。这种相互作用模式正适合将数据于分类讨论，根据不同的情况建立不同的体制，因此我们考虑门限自回归模型。

本文利用 1848 年至 1909 年数据进行建模。在建模的过程中，根据判断动态数据是否具有周期性质或者给出的数据是否具有非线性性质，建立了两种模型，即季节乘积模型 $ARMA(1,0)*(0,2)_{10}$ 和门限自回归模型 TAR(2,4,5,3)，并且分别对两种模型进行了一定的讨论。然后，利用 1910 年和 1911 年的数据进行预测值与真实值的对比，并使用所建立的模型对

作者简介：胡一睿（1989-），女，在读本科生，统计学. E-mail: doudou@mail.bnu.edu.cn

1912 年至 1916 年的数据进行预测。最后，以预测的平均误差作为不同模型之间的优良对比。

1 ARMA 模型

将预测指标随时间推移而形成的数据序列看作是一个随机序列，这组随机变量所具有的依存关系体现着原始数据在时间上的延续性。则称时间序列为 y_t 服从(p,q)阶自回归滑动平均混合模型。或者记为 $\varphi(B)\,y_t = \theta(B)\,\varepsilon_t$。

1.1 数据预处理

对貂年捕获量数据进行平方根变换与零均值化结合的预处理，达到平稳化的目的：

先对 X_t 进行变换：$W_t = \sqrt[4]{X_t}$，再将序列 $\{W_t\}$ 零均值化，使得到的序列 $\{Y_t\}$ 满足 $EY_t = 0$，即 $Y_t = W_t - EW_t = \sqrt[4]{X_t} - 14.93694$，如图一所示：

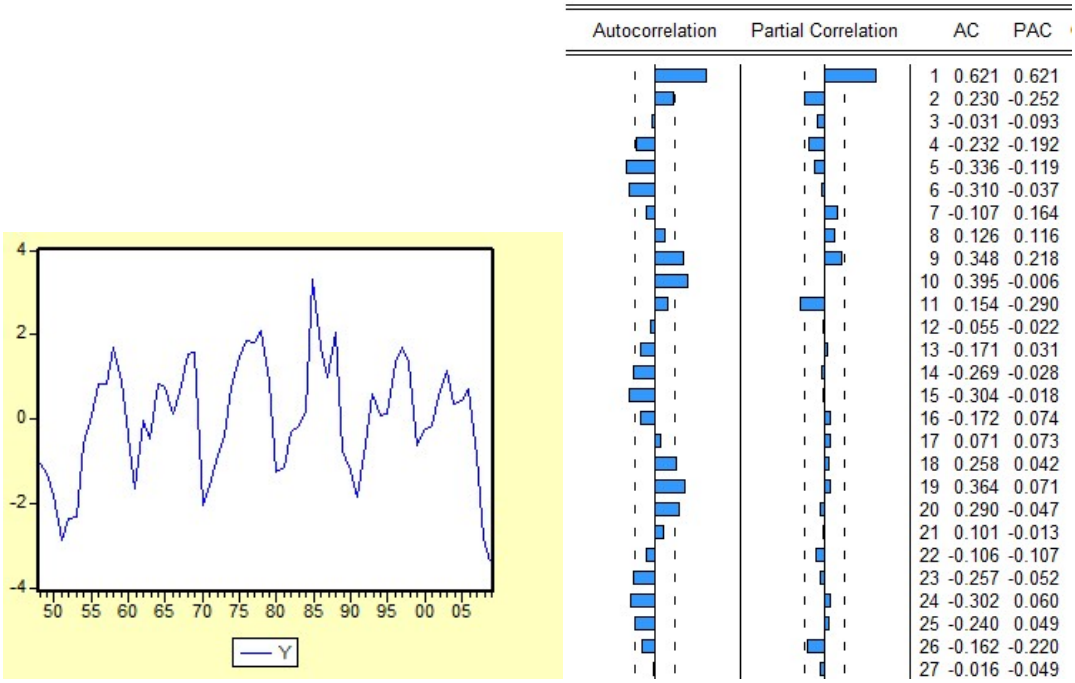

图 1 序列 $\{y_t\}$ 以及其相关函数

Fig.1 The correlation of $\{y_t\}$

1.2 建立季节乘积模型 ARMA(1,0)*(0,2)$_{10}$

观察序列 $\{y_t\}$ 以及其自相关函数图像，容易看出，序列 $\{y_t\}$ 有近似周期为 10。观察季节内外的自相关函数和偏相关函数（图一），发现季节内外的自相关函数和偏相关函数都无明显的结尾现象，类似于拖尾现象，并且是迅速衰减到零的，因此，假设模型为 ARMA(1,0)*(0,2)$_{10}$，利用 AIC 原则进行定阶 AIC 最小为 2.2746，此时，季节之内为 ARMA(1,0)模型，季节之间为 ARMA(0，2)模型。因此，估计建立模型为 ARMA(1,0)*(0,2)$_{10}$。

对模型进行单位根检验（上图下部分），可知模型的单位根均在单位圆以外，因此，该模型是稳定的。

因此，构建的模型为季节乘积模型 **ARMA(1,0)*(0,2)$_{10}$:**

(1-0.58)($\sqrt[4]{x_t}$ -14.94)=(1-0.60 B^{10})(1-1.82 B^{10}) ε_t, ε_t~N(0, 0.72^2).

1.3 ARMA 模型的预测结果

表 1 1910--1911 年的真实数据和拟合数据进行比较
Tab.1 The comparison of expected data in 1910 and 1911

年份	真实值	预测值	误差百分比
1910	21788	31709	45.53%
1911	33008	36418	10.33%

进一步，对 ARMA 模型进行五年的预测

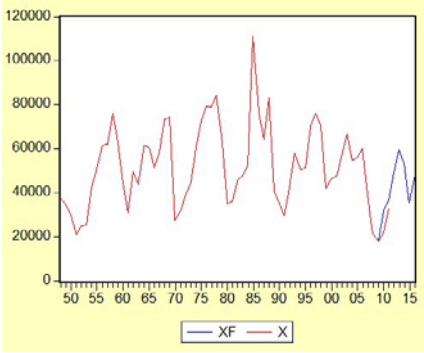

图 2 ARMA 模型五年预测
Fig.2 The prediction in the near 5 years

2 TAR 模型

门限自回归(TAR)模型是英籍华人H.TONG 博士于1978年首先提出来的，其基本思路是：在观测时序$\{X_t\}$的取值范围内引入l-1个门限值r_j，其中j=1,…,l-1。将时间轴分成l个区间，并用延迟步数d将$\{X_t\}$按$\{X_{t-d}\}$值的大小分配到不同的门限区间内，然后对不同区间内的$\{X_t\}$采用不同的AR模型来描述整个系统。

下面我们建立一个简单的门限自回归模型：

$$y_t = \varphi_0^{(j)} + \sum_{i=1}^{n_j} \varphi_i^{(j)} y_{t-i} + a_i^{(j)}, \quad r_{j-1} < y_{t-i} \le r_j, \quad j=1, 2, \ldots, 1$$

其中l是门限个数；d是延迟步数；r_j为门限值，j=1, …, l-1；对于每一个固定的j，$\alpha_i^{(j)}$是方差为σ_j^2的白噪声序列。

2.1 利用条件期望估计法确定门限个数

在二维坐标平面中，以X_{n-j}为横轴，$\tilde{E}(X_n \mid X_{n-j} \in R_l)$为纵轴，作点$(m_l, \tilde{E}(X_n \mid X_{n-j} \in R_l))$的图形，若点值图为线性分布，即对应平稳正态序列，则采用线性时序模型来描述；若点值图为非线性分布，则应采用非线性模型，特别地，当点值图呈分段线性分布时，应采取门限自回归模型，而且由点值图分段线性的个数可以确定门限自回归模型的门限区间数，由各线性段间的临界值确定门限值。[4] [5]

图 3 是横坐标依次为$X_{n-1}, X_{n-2}, X_{n-3}, X_{n-4}$,纵坐标依次为：

$\tilde{E}(X_n \mid X_{n-1} \in R_l), \tilde{E}(X_n \mid X_{n-2} \in R_l), \tilde{E}(X_n \mid X_{n-3} \in R_l), \tilde{E}(X_n \mid X_{n-4} \in R_l)$ 的四个点值图。

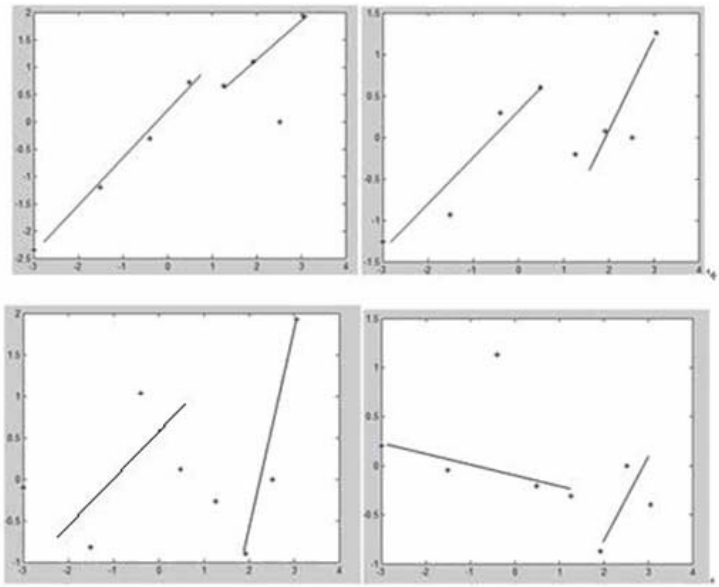

图 3 X_{n-j} 的点值图

Fig.3 The conditional expectation of X_{n-j}

只有点值图都可接受为直线时，才可以采用线性模型[4]，由上面的分析及点值图可确定：门限区间数由点值图分段线性的个数可以确定为 2，因此建立门限自回归模型 $TAR(2,4,5,3)$ 。

2.2 TAR 模型参数的估计

考虑体制规模的选取。模型自然分为捕获量增加期和捕获量减少期，又观察在 2.1 讨论中，线性模型的分类容易联想到建立体制为 2 的门限自回归模型。假设模型为 $TAR(2,p_1,p_2,d)$。下面将按照最小二乘法拟合参数，并且利用 AIC 原则确定 p_1,p_2,d 以及 r，其中 p_1 为体制一的自回归阶数， p_2 为体制二的自回归阶数， d 为延迟参数， r 为体制分类门限值。

首先固定 d ，让 r 在 $R=\{r_{30},r_{40},r_{50},r_{60},r_{70}\}$ 内变动，选取 $\tilde{r}$ 满足

$$AIC(d,\tilde{r})=\min_{r\in R}\{AIC(d,r)\}$$

再令 d 变动，选取 $\tilde{d}$ 满足

$$AIC(\tilde{d},\tilde{r})=\min_{d\in D}\{\frac{AIC(d,\tilde{r})}{n-n_d}\}$$

通过 AIC 准则得到最佳参数估计如下表：

表 2 1910--1911 年的真实数据和拟合数据进行比较
Tab.2 The comparison of expected data in 1910 and 1911

	D				
	1	2	3	4	5
$(\tilde{p}_1,\tilde{p}_2)$	（4，5）	（2，6）	（4，5）	（2，4）	（6，3）
n_d	4	2	4	4	6
$\tilde{r}$	51062	51062	45168	61745	51062
$AIC(\tilde{d},\tilde{r})$	0.0303	1.9303	-0.4923	0.5701	-0.2854

由上表可以看出，最小 AIC 为-0.4923，此时模型为$TAR(2,4,5,3)$，利用最小二乘法拟合的系数如下：$Y_t=\sqrt[4]{X_t}-14.93694$

$$Y_t=\begin{cases}0.7488y_{t-1}+0.1855y_{t-2}-0.4184y_{t-3}+0.0181y_{t-4}+\varepsilon_{1t}\\ x_{t-3}<45168,\varepsilon_{1t}\sim N(0,0.27554)\\ 0.6227y_{t-1}-0.2454y_{t-2}+0.2399y_{t-3}-0.2536y_{t-4}-0.4127y_{t-5}+\varepsilon_{2t}\\ x_{t-3}\geq 45168,\varepsilon_{2t}\sim N(0,1.0488)\end{cases}$$

利用 maple 求解两个自回归系数多项式的根，以此分别对两个 AR 模型系数进行单位根检验，可以知道系数方程的根在单位圆之外，因此，所建的模型是稳定的。

选取前 7 个自相关函数数据进行白噪声检验：71.43%的点在$(-1/\sqrt{58},1/\sqrt{58})$之间，由 71.43%>68%，由假设检验，不拒绝原假设，则得到的残差序列为白噪声。因此题述中的数据适用于$TAR(2,4,5,3)$模型。

2.3 TAR 模型的预测结果

表 3 1910--1911 年的真实数据和拟合数据进行比较
Tab.3 The comparison of expected data in 1910 and 1911

年份	真实值	预测值	$TAR(2,4,5,3)$ 误差百分比
1910	21788	22535	3.43%
1911	33008	32032	2.96%

从预测值和真值的比较观察，$TAR(2,4,5,3)$ 的预测是很接近的。从平均误差大小的角度而言，$TAR(2,4,5,3)$ 模型在近期预测较好。进一步，对 TAR 模型进行五年的预测，如下图所示：

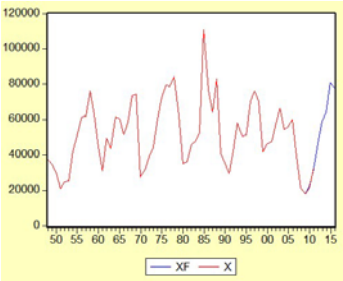

图 4 TAR 模型的五年预测值
Fig.4 The prediction of TAR in 5 years

3 ARMA 模型和 TAR 模型的比较

针对两个模型，分别从近期预测和长期预测来观察比较两个模型。

首先对两个模型近期预测值和真实值进行比较观察：

表 4 1910--1911 年的误差百分比比较
Tab.4 The comparison of MER in ARMA and TAR

	ARMA(1,0)*(0,2)10误差百分比	TAR(2,4,5,3)误差百分比
1910	45.53%	3.43%
1911	10.33%	2.96%
平均绝对误差	27.93%	3.20%

从预测值和真值的比较观察，ARMA(1,0)*(0,2)$_{10}$ 预测值容易偏大，而 TAR(2,4,5,3)的预测是很接近的。从平均误差大小的角度而言，TAR(2,4,5,3)模型在近期预测较好。

其次，从长期预测来看，预测图像如下所示：

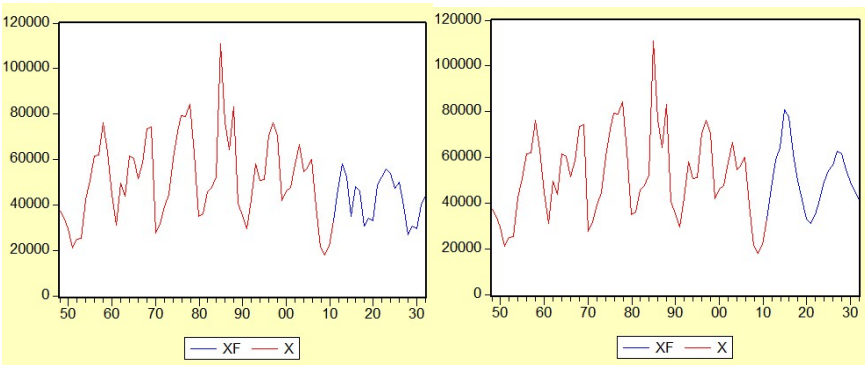

图 5 ARMA 和 TAR 模型的长期预测曲线

Fig.5 The prediction of ARMA and TAR in future

从长期预测的角度来看，从峰值角度而言，两者都不适用于预测，其峰值有减小的趋势。对于季节乘积模型，从一开始的近期预测误差难免偏大，后期峰值减弱现象更加明显，锯齿现象明显；并且，周期是人为给定的整值，从一定角度讲，有不合理的性质。然而，对于门限自回归模型，根据其动态数据的预测，自动有周期循环的趋势。长期预测虽然不能预测峰值，但是可以计算出动态数据形成的周期，在生态效应上，可以预测出貂捕获量的循环周期，这是符合生态模型的，关于周期在将门限自回归做更长期的观测很明显。

4 讨论

对于 ARMA 模型，我们知道该模型的建立是较为简便的，但是由 1910 年和 1911 年的预测数据和真实数据的比较可知，该模型对数据的预测有偏大的效应。而且，对于 1910 年的预测值而言，其误差是相当大的。而使该效应产生一个较关键的地方时周期的选取。在模型建立初期，我们有假设序列有近似的周期为 10，但是仔细观察序列，可以看出序列存在变周期的现象。数据的周期这一特性不是很明确，而采用季节乘积模型，在预测时采用滞后的周期数据会对预测的数据产生影响。因而，为了避免周期的影响，可以考虑摒弃对数据进行固定周期处理。

对于 TAR 模型，它考虑到变周期的不稳定性，只将数据变化看作是一种循环，从而避免了用固定周期数据延迟预测产生的影响；其次，针对动态数据的非线性性质，建立两种线性体制。同时，在该模型中，也考虑到生态效应在实际中的体现，故其不失为一个好的模型。然而，门限自回归模型的本质在于其为自回归模型，所以不能做长期预测。但在短期的预测中，拟合值与真实值相当接近，可以看出，其预测的误差是很小的。

由预报的原理，我们知道，随着时间的推移，预报的准确率将越来越小。但是我们做出几十年的趋势可以看出：TAR 预报曲线后面的稳定波动，这是有明显的生物意义的：若貂的捕获量维持在 1900 年时的捕获水平，则根据大自然和貂自身的不断调节适应，那么在 1950 年后，貂的捕获量就会呈现稳定的周期波动。

本文利用 1848 年至 1909 貂年捕获量数据进行建模。在建模的过程中，根据判断动态数据是否具有周期性质或者给出的数据是否具有非线性性质，建立了 ARMA 和 TAR 模型，通过分析可知，若数据具有明显的非线性，则采取门限自回归模型可以得到较好的预测效果。

[参考文献] (References)

[1] 范剑青 非线性时间序列 建模、预报及应用[M]. 北京 ：高等教育出版社，2005.
[2] 杜金观 时间序列分析[M]:建模与预报. 合肥 : 安徽教育出版社, 1991.
[3] 魏武雄 时间序列分析[M]: 单变量和多变量方法. 北京 : 中国人民大学出版社, 209.
[4] 安鸿志 时间序列分析[M]. 上海 : 华东师范大学出版社, 1992.3.
[5] 刘巍 计量经济学软件: Eviews 操作简明教程 [M].广州 ： 暨南大学出版社, 2009.
[6] 吴晓明；吴大文.自激励门限自回归模型研究与应用[J].沈阳航空工业学院学报，2001 年 04 期.