第32卷 第5期
2012年10月

桂林电子科技大学学报
Journal of Guilin University of Electronic Technology

Vol. 32, No. 5
Oct. 2012

基于ARIMA模型的时间序列建模算法和实证分析

赵肖肖，朱　宁，黄黎平

（桂林电子科技大学 数学与计算科学学院，广西 桂林　541004）

摘　要：通过对时间序列ARIMA模型建模方法的研究，将方差分析运用于时间序列建模，对季节数据做方差检验并确定周期。基于统计软件SAS分析ARIMA模型建模方法的具体算法，绘制详细的建模流程图。从模型的识别、参数估计、建模和预测等各方面介绍了模型建立和预测的全过程。利用SAS软件，结合引入的方差检验方法和算法流程对1990年1月至2010年12月的中国消费者价格指数季节性时间序列建立了乘积ARIMA模型，预测并分析了CPI的基本走势。

关键词：时间序列；ARIMA模型；季节模型；预测；方差分析；算法；CPI

中图分类号：O212.4　　**文献标识码**：A　　**文章编号**：1673-808X(2012)05-0410-06

Modeling algorithm and empirical analysis based on the time series of the ARIMA model

Zhao Xiaoxiao, Zhu Ning, Huang Liping

(School of Mathematics and Computational Science, GuiLin University of Electronic Technoligy, Guilin 541004, China)

Abstract: Through the study of time series ARIMA model modeling method, this paper applies variance analysis to time series modeling, after which carries out relevant variance tests on season datas and finally ascertains their cycle. Based on the detailed specific algorithm of statistical software SAS on analysing the ARIMA model modeling methods, as well as its specific steps on drafting particular flow chart. This paper elaborates the overall process of the model establishment and its forecast from those various aspects such as the model identification, the parameter estimation and the modeling establishment and its forecast. Finally, it uses the SAS software which combines with the incoming variance testing method and the algorithm process to establish the product ARIMA model on Chinese consumer price index of the seasonal time sequence from January 1990 to December 2010, forecast and analyze the basic trend of the CPI.

Key words: time series; ARIMA model; seasonal model; forecast; variance analysis; algorithm; CPI

时间序列的建模及预测在学术界和实际应用领域极为普遍，如城市交通量、人口增长量、股市价格、国民收入、太阳黑子数等预测问题[1]。自1970年Box和Jenkins的著作《时间序列分析、预测和控制》[2]问世以来，逐渐形成了一整套时间序列识别、估计、建模、预测及控制的理论和方法。以Box-Jenkins为代表的现代时间序列预测方法建立在随机过程理论基础上，具有结构简单、建模速度快、预测精度高等优点，并且解决了非平稳序列时间序列的处理问题，非常适合现实生活中各类随机性强的时间序列的分析和预测[2]。国内外已有许多学者对时间序列建模方法给予研究，袁振洲[3]在分析铁路货源流数据的内在规律及其时间序列特性的基础上，采用ARIMA模型对1989－1994年北京铁路局的货运煤炭总量进行

收稿日期：2012-04-15
基金项目：广西区"十一五"教学改革工程项目(GX06066)
通信作者：朱宁(1957－)，男，湖南宁乡人，副教授，研究方向为金融多元统计和数学建模。E-mail: znqx@guet.edu.cn
引文格式：赵肖肖，朱宁，黄黎平. 基于ARIMA模型的时间序列建模算法和实证分析[J]. 桂林电子科技大学学报，2012，32(5)：410-415.

了预测,通过对各种方法预测结果的综合比较和分析,发现 ARIMA 模型的预测效果最优。张利等[4]提出了一种对 ARIMA 模型改进的预测算法,并对短时交通流量时间序列建模,取得较为精确的预测结果。

时间序列是随时间改变而随机变化的序列,时间序列分析的目的是从中发现和揭示某一现象的发展变化规律,从而尽可能多地提取所需要的准确信息,并将这些知识和信息用于预测,以掌握和控制未来行为[5]。ARIMA 模型主要分为 3 种:自回归模型(AR 模型)、移动平均模型(MA 模型)和自回归移动平均模型(ARMA 模型)。求和自回归移动平均模型(简称 ARIMA 模型)主要是对非平稳序列建模,对非平稳序列进行平稳化后,即可按照 ARMA 模型的方法建立。

时间序列建模过程是动态过程,虽然明确 ARIMA 模型的建模的 3 个阶段[6],但如果没有精确的算法流程,难以快速有效地建立模型。同时,对于季节性数据,若用数理统计的方法对 ARIMA 建模过程所确定的周期给予检验,可以确保模型更为优越、预测结果更为准确。另外,实际中的数据大多具有异方差性,忽略异方差性无疑会丢失重要信息,影响预测结果。比如,文献[7-8]在研究中国 CPI 时间序列中没考虑异方差性。为此,给出方差检验方法确定季节数据的周期,将方差检验引入时间序列建模流程中,并基于统计软件 SAS 给出 ARIMA 模型建模方法的具体算法和流程图,利用 SAS 并结合引入的方差检验对中国 CPI 时间序列进行实证分析。

1　ARIMA 模型结构及方差分析

对时间序列 $\{x_t\}$,ARIMA(p,d,q)模型结构为

$$\Phi(B)\nabla^d x_t=\Theta(B)\varepsilon_t。$$

其中:p 为自回归模型的阶数;d 为差分阶数;q 为滑动平均模型的阶数;B 为延迟算子;$\Phi(B)=1-\varphi_1 B-\cdots-\varphi_p B^p$ 为自回归系数多项式;$\Theta(B)=1-\theta_1 B-\cdots-\theta_p B^p$ 为移动平滑系数多项式;$\{\varepsilon_t\}$ 为白噪声序列,$E(\varepsilon_t)=0$;$Var(\varepsilon_t)=\sigma^2<+\infty$;$\nabla^d=(1-B)^d$。

方差分析又称"变异数分析"或"F 检验",它是根据不同需要把某变量方差分解为不同的部分,比较它们之间的大小并用 F 检验进行显著性检验的方法。对于季节数据,如果将数据按周期排列,那么周期内的每个数据的特征都不相同,用方差分析可以检验。对于数据不满足均衡设计时,使用广义方差分析 GLM 检验[9]。方差分析的前提是数据服从正态分布,对于非正态数据,可以用非参数方法检验[9]。根据方差分析原理[10],将季节型时间序列 $\{x_t\}$ 按表 1 排列。

表 1　时间序列的设计表

Tab. 1　The design table of time series

V_1	V_2	…	V_m
x_1	x_2	…	x_m
x_{m+1}	x_{m+2}	…	x_{2m}
x_{2m+1}	x_{2m+2}	…	…

有 m 个总体 $V_i(i=1,2,\cdots,m)$(按表 1 设计,可称为变量),记 D_i 为第 i 个总体的方差。检验原假设 $H_0:D_1=D_2=\cdots=D_m$,备择假设 $H_1:D_1,D_2,\cdots,D_m$ 不全相等。如果检验结果 $p>0.05$,则接受原假设,说明变量个数不是时间序列的周期数。再增减变量个数重新设计序列;如果 $p<0.05$,则拒绝原假设,此时可以认为变量个数即为序列的周期。

2　基于 SAS 软件的 ARIMA 建模过程[11]

利用 SAS 软件可以对时间序列进行识别、参数估计、建模、预测,由 PROC ARIMA 过程执行的分析分为对应与 Box 和 Jenkins 描述的 3 个阶段,3 个阶段的 IDENTIFY、ESTIMATE、FORECAST 语句总结如下。

2.1　识别阶段

使用 IDENTIFY 语句来指定响应序列并且识别候选 ARIMA 模型。IDENTIFY 语句读入后面语句中使用到的时间序列,可对序列进行差分,然后计算出相关系数、逆自相关系数、偏自相关系数,此阶段的输出通常为一个或多个可拟合的 ARIMA 模型。

2.2　参数估计和诊断检验阶段

使用 ESTIMATE 语句能给出模型的参数估计,关于参数估计值的显著性检验可以指出模型中的一些项是否显著,拟合优度统计量可帮助比较该模型和其他模型的优劣。白噪声残差的检验可指明残差序列是否包含可被其他更复杂模型采用的额外信息。如果诊断检验表明模型不适用,可以尝试改变模型的

口径(即模型的阶数 p、q 的值)拟合另一个更优的模型。

2.3 预测阶段

使用 FORECAST 语句来预测时间序列的预报值,并对这些来自前面 ESTIMTE 语句生成的 ARIMA 模型的预测值产生置信区间。

3 ARIMA 模型的建模算法及流程图

ARIMA 模型建立的基础是时间序列具有平稳性和方差齐性,因为如果用非平稳序列来建立模型,就会出现虚假回归问题。即尽管基本序列不存在任何关系,也会得到回归模型;如果序列存在异方差性,则会影响模型的拟合效果。

ARIMA 模型的建模思想可描述为:对一列原始数据,首先应当先判断该序列是否为平稳序列,若不是,则用平滑法或差分法对原始数据进行平稳化,再对序列进行建模。建模算法如 Step1~11,流程图见图 1。

Step1:获得观测值序列;

Step2:做原序列时序图,若序列有趋势,前进 Step3;若没有趋势,前进 Step5;

Step3:对序列做一阶差分消去趋势,前进 Step4;

Step4:对差分序列做时序图;若还有趋势,返回 Step3;若无趋势,前进 Step5;

Step5:对序列方差齐性检验,若有异方差性,前进 Step6;若方差齐性,前进 Step7;

Step6:对原始数据取对数变换,返回 Step2;

Step7:模型识别(IDENTIFY),确定延迟阶数 p、q,前进 Step8;

Step8:参数估计和模型诊断检验(ESTIMATE),前进 Step9;

Step9:预测(FORECAST);若是对数数据,前进 Step10;若是原始数据,前进 Step11;

Step10:将预测值、95%的置信上下限进行指数化,变换为原始数据单位,前进 Step11;

Step11:做真实值和预测值的时序图,结束。

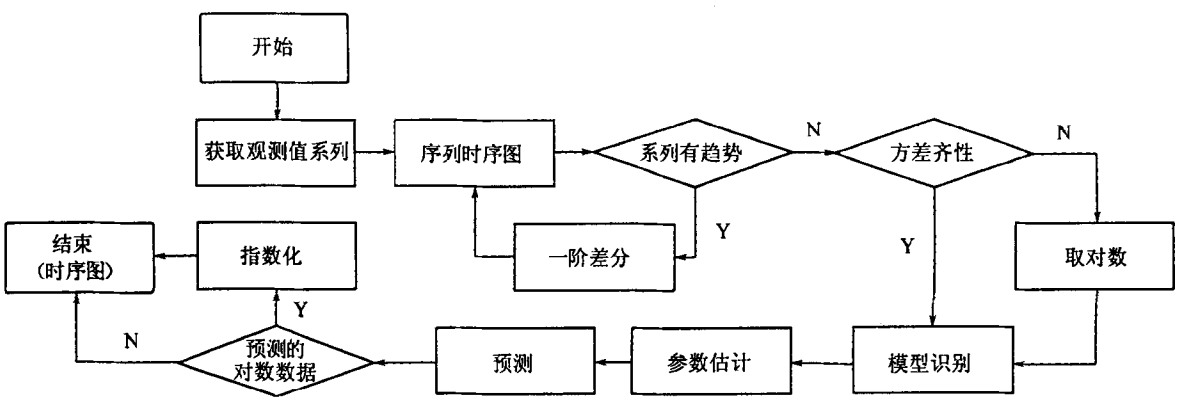

图 1 ARIMA 模型建模算法流程图

Fig. 1 The flow chart of ARIMA modeling algorithm

4 实例研究

根据 ARIMA 模型的建模算法和流程图,分析 1990 年 1 月至 2010 年 12 月的中国消费者价格指数时间序列的建模。

4.1 识别阶段

4.1.1 获得观测值序列

1990 年 1 月至 2011 年 7 月的中国 CPI 月数据(数据来源:凤凰网 www. ifeng. com),其中采用 1990 年 1 月至 2010 年 12 月的数据拟合模型,并预测 2011 年 1~7 月的 CPI 数据。

4.1.2 判断模型的平稳性

时间序列模型是建立在随机序列平稳性假设的基础上的,因此,序列的平稳性是建模的重要前提。通过原始序列的时序图可以判断序列是否有趋势,由 PROC GPLOT 过程可得该序列的时序图如图 2 所示,序列有明显的下降趋势。需要对序列做差分从而消去趋势,因为建立时间序列模型的出发点是以统计独立的白噪声作为输入极力源,它通过一贯线性的动态系统输出所需要的时间序列模型[12]。

4.1.3　对原序列平稳化

对序列进行一阶差分,虽然一阶差分消去了趋势项,但由于序列波动性在 1990－1996 年和 2005－2010 年间较 1997－2004 年间大,表明该序列具有异方差性,还不能进行模型的识别,因为 ARIMA 模型是建立在序列方差齐性基础上的。因此,需要对时间序列进行方差齐性变换。

4.1.4　序列方差齐性变换

通常假定已知异方差函数具体形式,则需对原始序列进行方差齐性变换。方差齐性变换在理论上有明确的意义,变换函数是由异方差函数 $Var(x_t)=\sigma_t^2$ 在均值 μ_t 附近作一阶泰勒展开而得到的,即对原始序列取对数,对数变换在大多数时间序列分析时被普遍采用。于是首先对原始序列取对数,进行方差齐性变换。由图 3 所示的取对数数据一阶差分时序图可以看出,原始数据经过对数变换后已基本接近方差齐性,从而可以对取对数数据进行模型识别。

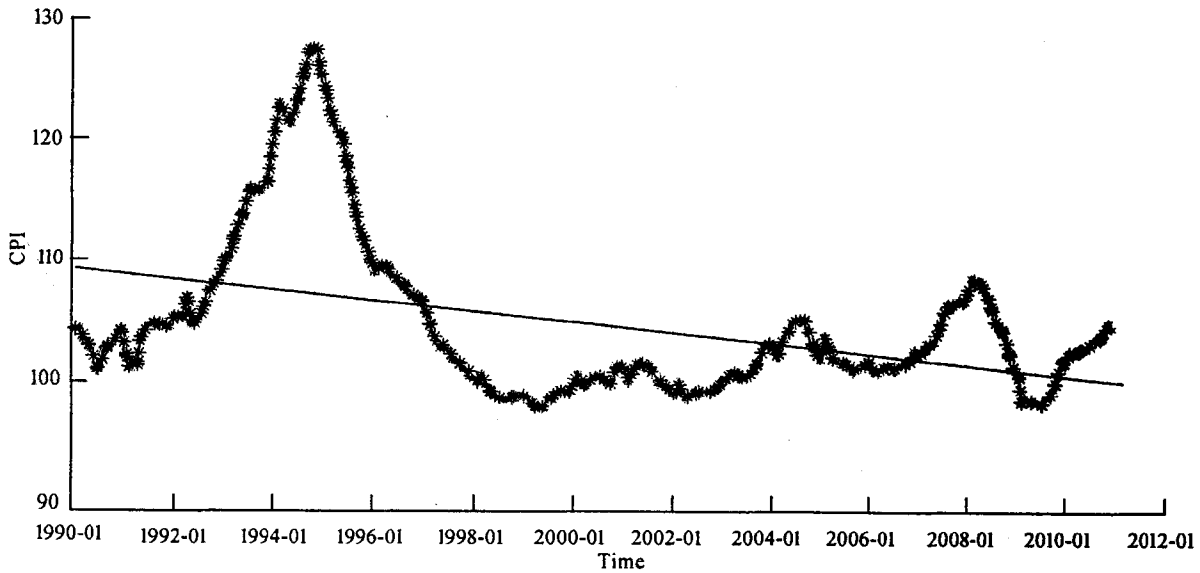

图 2　1990－2010 年中国 CPI 月数据时序图

Fig. 2　The plot of Chinese CPI sequence between 1990 and 2010

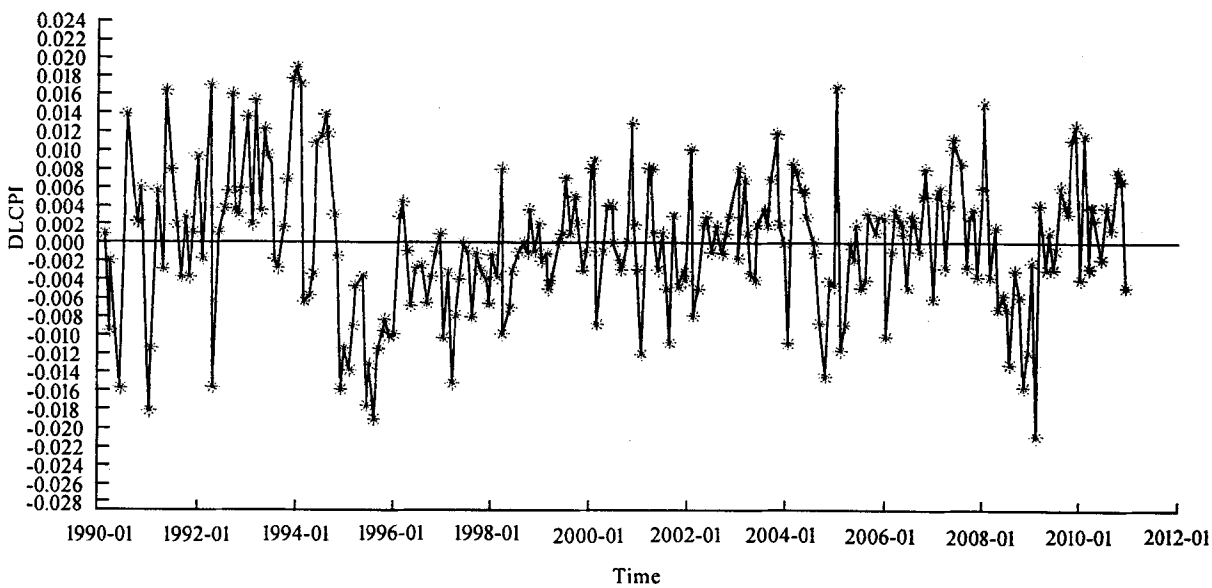

图 3　取对数数据一阶差分数据的时序图

Fig. 3　The plot of an order difference series on logarithmic data

4.1.5 模型识别

直接调用 PROC ARIMA 过程的 IDENTIFY 语句实现对所选差分时滞的检验，目的是确定所选差分序列时滞情况下的 ARIMA 模型的阶数 p、q 值。

考察取对数一阶差分序列自相关图(见图 4)，自相关系数在延迟 12 阶，明显增大，于是将数据按 12 个变量顺序排列设计，进行方差检验。方差检验结果见表 2，F 值为 171.38，$P<0.000\ 1$，从而确定序列周期为 12。进行 12 步的差分来消去季节效应，再对序列进行白噪声检验，延迟 12 阶、24 阶的 P 值都远远小于 0.05，故消去季节效应后的序列为非白噪声序列，说明该序列有信息可以提取，即可以对此时间序列进行建模。

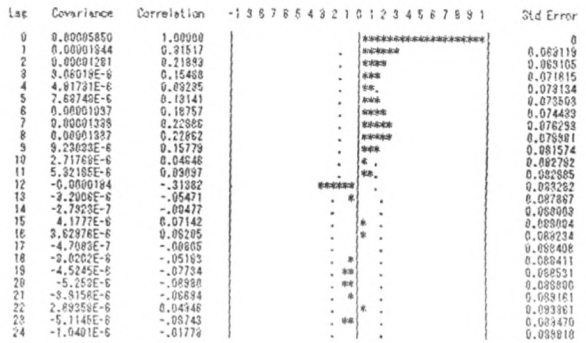

图 4 取对数数据一阶差分的自相关图

Fig. 4 The autocorrelation plot of an order difference series on logarithmic data

表 2 方差分析结果

Tab. 2 The results of variance analysis

变异来源	离差平方和	均方	F 值	P 值
组间	10176.557 7	508.827 88	171.38	<0.000 1
组内	685.827 5	2.968 95	—	—
总方差	10 862.385 2	—	—	—

4.2 参数估计和诊断检验阶段

采用非线性最小二乘(NLS)估计模型的参数。由 AR 模型具有拖尾的自相关系数。截尾的偏自相关系数；MA 模型具有截尾的自相关系数，拖尾的偏自相关系数。根据自相关图和偏自相关图，初步识别取对数模型为 $p=1$、12 和 $q=1$、12 的乘积季节模型[6]。但是到目前为止该类模型中对 p、q 的确定尚无定论[5]，一般情况下根据自相关系数和偏自相关系数衰减幅度确定 p、q，再根据 AIC 准则和 SBC 准则，可以有效地弥补根据自相关图和偏自相关图定阶的主观性，取得在所有检验的模型中使得 AIC 或 SBC 函数达到最小的相对最优模型[6]。但是不能只考虑这 2 个准则，而忽视拟合模型残差是否还包含没有提取完的信息。各模型的检验见表 3。

表 3 模型的检验信息

Tab. 3 The model test information

	$p=1$、12 $q=1$、12	$p=2$、12 $q=1$、12	$p=1$、12 $q=2$、12	$p=2$、12 $q=2$、12
标准误估计	0.008 353	0.008 535	0.008 539	0.008 665
AIC	−1 605.1	−1 594.74	−1 594.57	−1 587.56
SBC	−1 591.19	−1 580.84	−1 580.66	−1 573.66
残差白噪声(延迟 6 阶 p 值)	0.073 5 (白噪声)	0.005 5 (非白噪声)	0.004 8 (非白噪声)	0.000 9 (非白噪声)

分析以上结果，最终确定我国近 20 年消费者价格指数时间序列乘积季节模型为 ARIMA(1,1,1)×$(1,1,1)_{12}$，参数估计和检验结果见表 4。

表 4 模型的参数估计

Tab. 4 The model parameter estimation

待估参数	估计值	标准误	t 统计量	P 值
θ_1	0.715 94	0.086 7 8	8.25	<0.000 1
θ_{12}	0.662 84	0.060 2 8	11.00	<0.000 1
φ_1	0.904 65	0.539 8	16.76	<0.000 1
φ_{12}	−0.368 13	0.076 0 1	−4.84	<0.000 1

显然各参数都是显著的，拟合模型如下：

$$(1-B)(1-B^{12})\log(x_t)=\frac{(1-0.71594B)(1-0.66284B^{12})}{(1-0.90465B)(1-0.36813B^{12})}\varepsilon_t。$$

其中：B 为延迟算子；ε_t 为随机干扰序列。

4.3 预测阶段

在预测阶段必须明确一点，FORECAST 语句预测的是处理后数据的预测值。本研究预测的是对数数据值，要想得到原始数据的预测值，必须要将对数值的预测变换回原来的测量单位。CPI 预测的结果见表 5，预测曲线见图 5。这里误差率为

$$e_t=\frac{|x_t-\hat{x}_t|}{x_t}\times 100\%。$$

其中：e_t 为误差率；x_t 为真实值；$\hat{x}_t$ 为预测值。

表 5　2011 年 1～7 月 CPI 预测值

Tab. 5　The CPI prediction from January to July in 2011

	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07
真实值	104.90	104.90	105.40	105.30	106.50	106.40	106.50
预测值	104.90	104.61	104.97	105.08	105.24	105.36	105.46
误差率/%	0	0.28	0.41	0.79	1.19	0.98	0.98

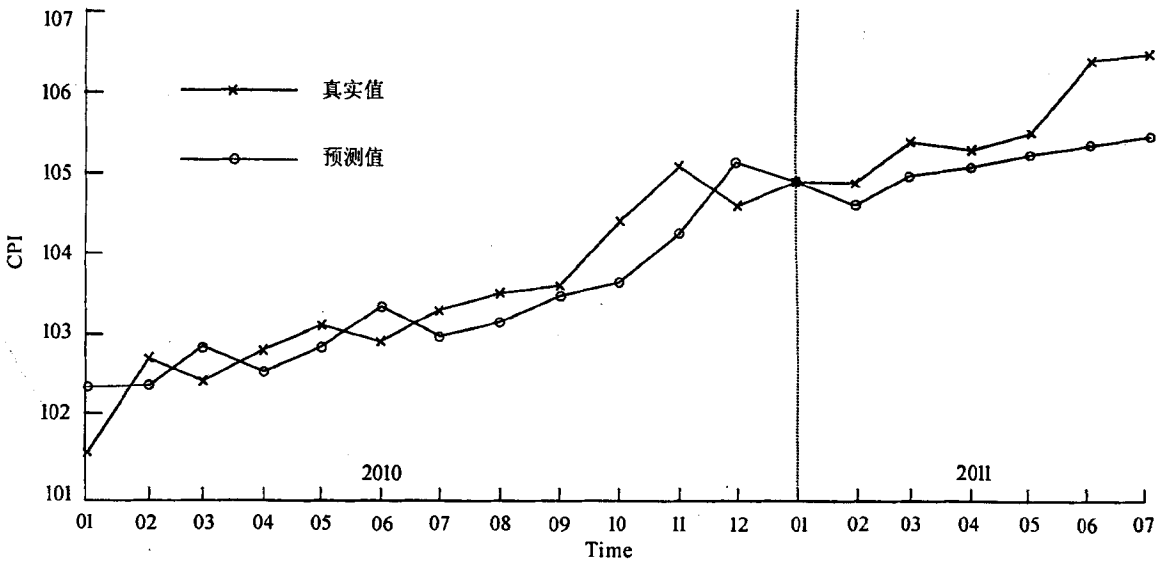

图 5　2010-01—2011-07 CPI 预测曲线

Fig. 5　The CPI prediction curve in 2010-01—2011-07

5　结束语

在解决一个实际问题时，确定一个清晰的算法和流程图对解决问题有很大的帮助。本研究给出的时间序列 ARIMA 模型的建模算法和流程图基于 SAS 软件而作，同时，利用方差分析来检验季节模型的周期，增强建模方法的效率。依据算法和流程图对我国 CPI 时间序列建立了一个较优的模型。分析结果可以看出，随着预测步长的增加，预测误差也变大，该模型只考虑了时间序列本身的特性，而没有考虑其他一些不确定因素对消费者价格指数的影响，模型作短期预测能够保证预测的精确性。基于 SAS 系统对近 20 年来中国消费者价格指数月数据建立了季节 ARIMA$(1,1,1)\times(1,1,1)_{12}$模型，该模型可以反映我国 CPI 的变化规律，对有关部门制定一些相关政策具有极大的参考价值。

参考文献：

[1] 霍俊. 实用预测学[M]. 北京：中国发明创造者基金会，中国预测研究会，1984：31-87.

[2] Box G E, Jenkins G M. 时间序列分析预测与控制[M]. 北京：中国统计出版社，2003：25-90.

[3] 袁振洲. 应用自回归积分移动平均法预测铁路货源货流发展趋势[J]. 铁道学报，1996(18)：52-56.

[4] 张利，李星毅，施化吉. 一种基于 ARIMA 模型的短时交通流量改进预测算法[C]//2007 第三届中国智能交通年会论文集. 南京：东南大学出版社，2007.

[5] 汤岩. 时间序列分析的研究与应用[D]. 哈尔滨：东北农业大学，2007.

[6] 王燕. 应用时间序列分析[M]. 北京：中国人民大学出版社，2005：68-87.

[7] 谢佳利，杨善朝，梁鑫. 我国 CPI 时间序列预测模型的比较及实证检验[J]. 统计与决策，2008(9)：4-6.

[8] 汪淼，郑舒婷. 基于 ARIMA 模型的中国消费者价格指数时间序列分析[J]. 辽宁工程技术大学学报，2010，29：130-132.

[9] 黄燕，吴平等. SAS 统计分析及应用[M]. 北京：机械工业出版社，2006：103-118.

[10] 高惠璇. 应用多元统计分析[M]. 北京：北京大学出版社，2005：66-92.

[11] 高惠璇. SAS 系统 SAS/ETS 软件使用手册[M]. 北京：中国统计出版社，1998：65-72.

[12] 朱宁，徐标，仝殿波. 上证指数的时间序列模型[J]. 桂林电子工业学院学报，2006，26(2)：124-128.

编辑：梁王欢