【检验技术与方法研究】

ARIMA 乘法季节模型的 R 软件实现

李亚伟 刘玲 宋士勋 路凤

摘要: 目的 探讨 ARIMA 乘法季节模型的 R 软件实现方法,为模型的利用提供方法参考。**方法** 利用美国芝加哥市 1987—2000 年大气污染物臭氧(O_3)浓度数据建立 ARIMA 乘法季节模型,并进行预测,比较预测值和观察值的差异。**结果** ARIMA 乘法季节模型在 R 软件中方便实现,模型预测值和观察值的平均相对误差为 5.6%。**结论** R 软件有相对丰富的软件包可以实现 ARIMA 乘法季节模型,使用者可以方便快捷地实现分析需求。

关键词: ARIMA 乘法季节模型; 大气污染物; 预测; R 软件实现

中图分类号: R122

DOI: 10.13421/j.cnki.hjwsxzz.2018.04.013

Application and R Software Implementation of Multiplicative Seasonal ARIMA Model

LI Yawei, LIU Ling, SONG Shixun, LU Feng

Abstract: Objectives To explore the R software implementation of multiplicative seasonal ARIMA model and to provide a method for its utilization. **Methods** The ARIMA model was established on account of the atmospheric pollutant ozone (O_3) concentration from 1987 to 2000 in Chicago, USA, and the difference between predicted value and observed value was compared. **Results** ARIMA model was implemented conveniently in R software. and the average relative error between the predicted and observed values was 5.6%. **Conclusions** R software had a relatively abundant useful software packages for fitting the multiplicative seasonal ARIMA model, and users could complete the analysis conveniently and quickly.

Key words: multiplicative seasonal ARIMA model, air pollutant, prediction, R software implementation

时间序列数据可依据变量自身的变化规律,利用外推机制描述序列变化,建立模型并对未来趋势进行预测[1]。ARIMA 乘法季节模型是研究这种时间序列数据的经典方法[2],R 软件作为一种自由、开源和免费的软件,应用越来越广泛,利用 R 软件实现 ARIMA 乘法季节模型的报道多为应用性研究,缺乏详细操作步骤和流程,本文以大气污染物臭氧(O_3)浓度数据为例,就 ARIMA 乘法季节模型的 R 软件实现做一介绍。

1 ARIMA 乘法季节模型简介

ARIMA 乘法季节模型是 ARIMA 模型的一种特殊形式。ARIMA 模型全称为自回归移动平均混合模型(Autoregressive Integrated Moving Average Models)。包括自回归模型(autoregressive model, AR)、差分(Integrated, I)、移动平均模型(moving average model, MA) 3 个部分,是美国统计学家 Geogre E. P. Box 和英国统计学家 GunlymM. Jenkins 创立的时间序列分析方法中的基本模型之一[2]。

作者简介: 李亚伟,助理研究员,从事环境流行病学方向研究
作者单位: 中国疾病预防控制中心环境与健康相关产品安全所
联系方式: 北京市西城区南纬路 29 号; 邮编: 100050; Email: liyawei@nieh.chinacdc.cn
通信作者: 路凤,副研究员,从事环境流行病学研究,Email: lufeng@nieh.chinacdc.cn

1.1 AR、MA、ARMA(p,q)、ARIMA(p,d,q)、ARIMA(p,d,q)×(P,D,Q)$_s$ 模型[2-3]

对于时间序列数据 x_t,AR 模型具有如下结构

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_{pxt-p} + u_t$$

该模型表示,t 时刻的观察值与前 p 个时刻的观察值呈线性关系,称为 p 阶自回归,φ 为自回归系数,记为 AR(p)。u_t 表示残差序列,又称白噪声(white noise)序列。用后移动算子 B 表示为 $\varphi(B)x_t = u_t$,其中 $\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$; MA 模型具有如下结构

$$x_t = u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \cdots - \theta_{qut-q},$$

表示 x_t 在 t 时刻的取值与 q 个随机扰动呈线性关系,称为 q 阶移动平均模型,θ 为滑动平均系数,记为 MA(q),用后移动算子 B 表示为 $x_t = \theta(B)u_t$,其中 $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$; 对于 AR(p)模型,若 u_t 不是白噪声,表现为 MA(q)的形式,则为 ARMA(p,q)模型,即: $\varphi(B)x_t = \theta(B)u_t$。ARMA(p,q)模型要求做预测的时间序列为平稳序列,图形显示所有观察值围绕某一水平直线上下随机波动,若序列不平稳,需经过差分变换,使其平稳,即 ARIMA(p,d,q)模型,d 为差分次数。

某些以月和季度为单位的数据,如大气污染物的月均浓度,常具有明显的季节特征,季节模型的时间单位为相应的周期 S,表示为 ARIMA(P,D,Q)$_s$,即 $\Phi(B^S)W_t = \Theta(B^S)u_t$,P、D、Q 分别表示以 s 为间距的自回归、差分、移动平均阶数,令 $W_t = \nabla^d \nabla_S^D x_t$,表示经过差分和季节差分得到的序列,其中 $\Phi(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS}$,$\Theta(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \cdots - \Theta_Q B^{QS}$,上述全部模型综合在一起即为 ARIMA 乘法季节模型,表示为:

$$\varphi_p(B)\Phi_P(B^S)W_t = \theta_q(B)\Theta_Q(B^S)u_t$$

记为: ARIMA(p,d,q)×(P,D,Q)$_s$,模型需要对 p、d、q、P、D、Q 进行定阶,从而进行估计。

1.2 模型的分析步骤[2,4]

1.2.1 序列平稳化和纯随机性检验 通过序列图、自相关系数函数图(auto correlation function,ACF)、偏自相关系数函数图(partial autocorrelation function,PACF)等分析序列平稳性,查看是否有趋势性、季节性等特征,并对序列进行 ADF 检验,考察平稳性。若序列不平稳,需要进行差分和(或)季节差分,将序列转换成平稳序列; 利用 LB(Ljung-Box)统计量对序列进行白噪声检验: 若为非白噪声,说明序列间存在相关关系,可以进行分析。

1.2.2 模型识别与定阶 根据差分次数、自相关函数、偏自相关函数等初步确定 p、d、q、P、D、Q 的阶数,基本原则是: ACF 图拖尾,PACF 图 p 阶截尾,选择 AR(p)模型; ACF 图 q 截尾,PACF 图拖尾,选择 MA(q)模型; ACF 图、PACF 图均拖尾选择 ARMA(p,q)模型。

1.2.3 参数估计与模型诊断 利用极大似然估计、最小二乘估计等方法,估计模型参数,参数估计后,对模型的残差进行白噪声检验,判断模型的适用性。依据赤池信息准则(AIC)、贝叶斯信息准则(BIC)等方法调整模型阶数,对模型进行比较和优化,选取较为简洁的最佳模型。

1.2.4 预测应用 用选定的模型对序列进行预测。

2 材料与方法

利用美国芝加哥市大气污染物臭氧(O_3)数据为例,进行模拟,数据来源于 R 自带的程序包,数据库汇总了 1987—2000 年芝加哥市逐日 O_3 浓度,将其整理成月均浓度数据分析(表 1)。其中,1987 年 1 月—1999 年 12 月的数据用于建立模型,所建模型用来预测 2000 年(1—12)月的月均浓度,将预测值和实际观察值进行比较,计算相对误差,对模型进行评估。

表 1 芝加哥市臭氧浓度数据(部分)

年份/年	月份/月	臭氧浓度/(μg/m^3)
1987	1	7.7
1987	2	11.9
1987	3	23.4
…	…	…
2000	10	16.4
2000	11	8
2000	12	10.4

3 R 软件实现[2,5]

3.1 序列平稳性和纯随机性检验

```
> chi <-read.csv("chi.csv", header = TRUE, sep = ',') #导入数据
> o3 <-ts(chi $ o3, start = c(1987,1), frequency = 12) #将臭氧浓度转为时间序列变量,频率为 12,即为月度数据。
> plot(o3, type = 'o', ylab = "臭氧浓度", xlab = "年份", main = "原始序列图") #绘制序列图
> acf(as.vector(o3), lag = 36, main = "原始序列") #绘制序列自相关图(图 1)。
```

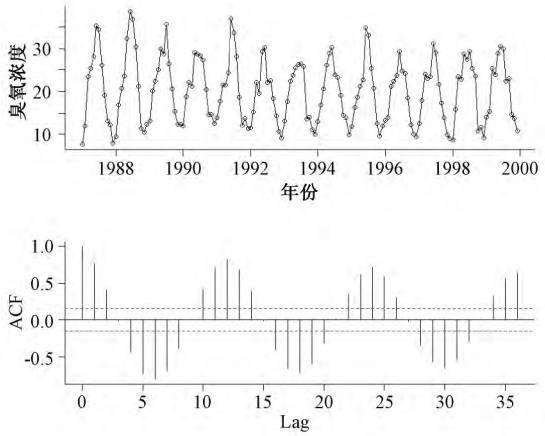

图 1 原始序列图

图 1 显示，原始序列均值稳定在 20 左右，没有明显的上升、下降趋势，序列有较强的季节特征，表现为夏季月份偏高，冬季月份偏低，自相关函数在滞后 12 阶，24 阶，36 阶…上有强相关性。考虑进行季节差分，消除季节趋势，使用序列变为平稳序列。

> o3_D1 = diff(o3 ,lag = 12) #进行季节差分，形成一次季节差分序列

> library('forecast') #载入"forecast"包

> tsdisplay(o3_D1 ,lag = 48, main = '一次季节差分序列') #绘制 o3_D1 序列图、自相关函数图、偏自相关函数图(图 2) 。

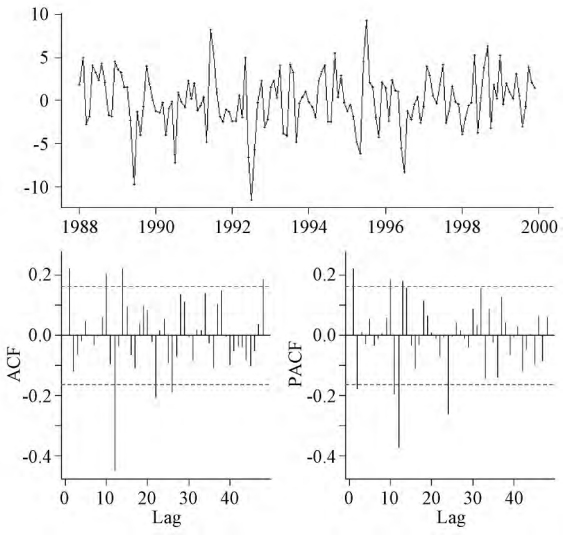

图 2 一次季节差分序列

图 2 序列图、自相关函数图显示经过季节差分，原始序列明显的季节趋势已经基本消除，可以认为 o3_D1 序列为平稳序列。

> library(tseries) #载入"tseries"程序包

> adf. test(o3_D1) #单位根检验

$P < 0.01$，拒绝序列为非平稳序列的原假设，可以认为一次季节差分后的序列为平稳序列。

> Box. test(o3 , lag = 6 , type = " Ljung-Box") ; Box. test(o3 , lag = 12 , type = " Ljung-Box") #纯随机性检验

LB 统计量结果结果显示，滞后 6 阶和滞后 12 阶，P 值均小于 0.01，在 $\alpha = 0.05$ 水平，拒绝序列为纯随机的原假设，可以认为原始序列为非白噪声序列，可以进行后续分析

3.2 模型定阶

根据图 2，季节差分后的序列，ACF 图在 1 阶、12 阶后，PACF 图在 2 阶、24 阶后快速衰减，表现出截尾特征，初步选定 p = 2、d = 0、q = 1、P = 2、D = 1、Q = 1。

3.3 模型拟合与优化

> fit1 < -arima(o3 , order = c(2 ,0 ,1) , seasonal = list(order = c(2 ,1 ,1) , period = 12) , method = 'ML') ##拟合模型，选用极大似然估计的方法

> acf(as. vector(fit1 $ residuals) , lag = 36 , main = "残差序列") ;

pacf(as. vector(fit1 $ residuals) , lag = 36) #绘制残差的 ACF、PACF 图(图 3) 。

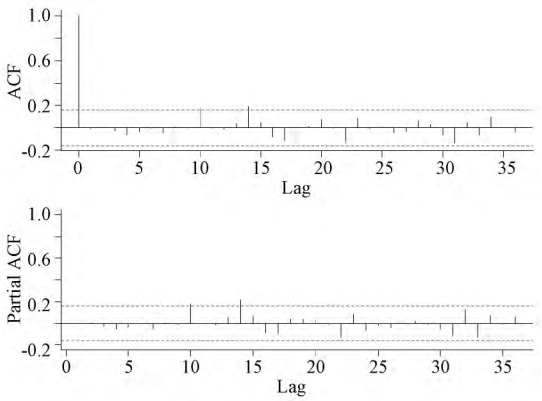

图 3 残差序列

图 3 显示残差的绝大部分自相关函数、偏自相关函数已经落入 2 倍标准差之内，可以认为残差序列为白噪声。

> Box. test (fit1 $ residuals , lag = 6 , type = " Ljung-Box") ; Box. test(fit1 $ residuals , lag = 12 , type = " Ljung-Box") #残差的白噪声检验

LB 统计量结果结果显示，滞后 6 阶和滞后 12 阶，P 值均大于 0.05，在 $\alpha = 0.05$ 水平，可以认为残差序列为白噪声序列，模型 1 显著有效。

> fit2 <- arima(o3, order = c(0,0,1) , seasonal = list(order = c(0,1,1) , period = 12) , method = 'ML') #经过多次试探性降阶尝试,拟合 AIC 值较小、更简洁的模型 2

AIC(fit1, fit2) #比较两个模型的 AIC 值

结果显示 AIC2(701.9) < AIC1(707.3) ,选择模型 2 作为最终模型。

3.4 模型预测

> o3. fore = forecast(fit2, h = 12, level = 0.95) #利用模型 2 进行一个周期,即 2000 年预测

> summary(o3. fore) #查看预测结果

> plot(o3. fore) #绘制预测图(图 4) 。

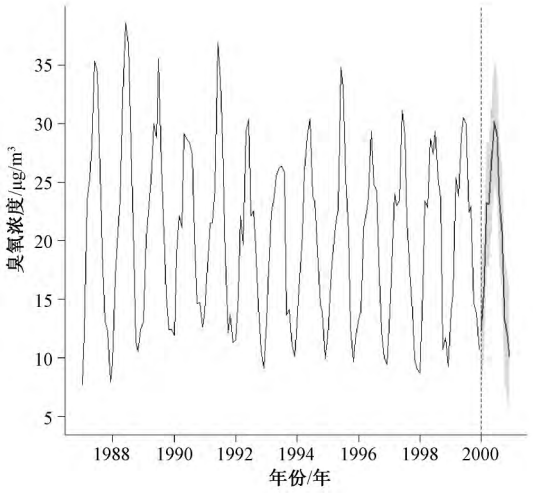

图 4 2000 年预测图

4 结果

2000 年臭氧浓度的预测值和观察值比较见表 2,结果显示平均相对误差为 5.6%。

表 2 2000 年 1—12 月臭氧浓度预测值和观测值比较

年份	月份	观测值	预测值	绝对误差	相对误差/%
2000	1	12.8	12.3	-0.5	-3.9
2000	2	17.6	16	-1.6	-9.1
2000	3	20.9	23.2	2.3	11
2000	4	21.4	23.1	1.7	7.9
2000	5	23.9	26.9	3	12.6
2000	6	25.3	30.1	4.8	19
2000	7	28.2	28.9	0.7	2.5
2000	8	25.3	23.8	-1.5	-5.9
2000	9	19.4	20.8	1.4	7.2
2000	10	16.4	13.2	-3.2	-19.5
2000	11	8	11.9	3.9	48.8
2000	12	10.4	10	-0.4	-3.8
平均相对误差					5.6

5 小结

本文对 ARIMA 乘法季节模型做了简要的介绍,以有季节特征的臭氧浓度数据为例,给出了用 R 软件实现模型拟合、检验、图形绘制以及预测的详细过程。R 软件由于其较大的灵活性,近年来在数据分析、作图等领域应用非常广泛,不同的统计方法通常都能找到相应的软件包,便于使用者使用,本文使用了 forecast、tseries 两个软件包。ARIMA 模型通过调整相关参数,也可以实现以日、周为单位的数据分析[6]。除大气污染物的数据外,类似具有相关特征的数据也可采用此方法进行预测,如医院的门诊量[7]、日死亡数、传染病的发病数据[8]等,在实际工作中使用者应根据数据特征进行合理分析。

参考文献

[1] 温亮,张秀山,李承毅,等. 季节分解法和 ARIMA 法预测乌鲁木齐市肺结核发病趋势效果分析[J]. 军事医学,2017,41(4):287-290. (In English: Wen L, Zhang XS, Li CY, et al. Seasonal decomposition and ARIMA methods in prediction of tuberculosis incidence in Urumqi, China[J]. Mil Med Sci, 2017, 41(4): 287-290.)

[2] 王燕. 应用时间序列分析[M]. 3 版. 北京:中国人民大学出版社,2012. (In English: Wang Y. Applied Time Series Analysis[M]. 3rd ed. Beijing: China Renmin University Press, 2012.)

[3] 张文彤,董伟. SPSS 统计分析高级教程[M]. 2 版. 北京:高等教育出版社,2013.

[4] 胡跃华,廖家强,冯国双,等. 自回归移动平均模型在全国手足口病疫情预测中的应用[J]. 疾病监测,2014,29(10):827-832. (In English: Hu YH, Liao JQ, Feng GS, et al. Application of multiple seasonal autoregressive integrated moving average model in prediction of incidence of hand foot and mouth disease in China[J]. Dis Surve, 2014, 29(10): 827-832.)

[5] Cryer JD, Chan KS. 时间序列分析及应用:R 语言[M]. 北京:机械工业出版社,2011. (In English: Cryer JD, Chan KS. Time Series Analysis with Applications in R[M]. Beijing: China Machine Press, 2011.)

[6] 牟敬锋,赵星,樊静洁,等. 基于 ARIMA 模型的深圳市空气质量指数时间序列预测研究[J]. 环境卫生学杂志,2017,7(2):102-107,117. (In English: Mou JF, Zhao X, Fan JJ, et al. Time series prediction of AQI in Shenzhen based on ARIMA model[J]. J Environ Hyg, 2017, 7(2): 102-107, 117.)

[7] 杨帆,秦银河,刘丽华. ARIMA 模型在门诊人次预测中的应用[J]. 中华医院管理杂志,2009,25(1):28-31. (In English: Yang F, Qin YH, Liu LH. Application of ARMA Model in prediction of outpatient headcount[J]. Chin J Hosp Admin, 2009, 25(1): 28-31.)

[8] 王建书,刘强,覃江纯,等. 基于 ARIMA 乘积季节模型的苏州市介水传染病发病预测研究[J]. 环境卫生学杂志,2017,7(6):417 - 420. (In English: Wang JS, Liu Q, Qin JC, et al. Prediction of incidence for water-borne diseases on a multiple seasonal ARIMA model in Suzhou [J]. J Environ Hyg, 2017, 7(6): 417 - 420.)

(责任编辑:鲁波)

(上接第 344 页)

[3] 张潜,高舸,王炼,等. 自动固相微萃取—气相色谱/质谱法快速筛查饮水中 45 种挥发性和半挥发性有机物[J]. 实用预防医学,2016,23(3):275 - 279. (In English: Zhang Q, Gao G, Wang L, et al. Automated solid phase micro-extraction coupled with gas chromatography-mass spectrometry for rapid analysis of 45 volatile and semivolatile organic compounds in drinking water [J]. Pract Prev Med, 2016, 23(3): 275 - 279.)

[4] 王艳丽,周阳. 固相微萃取—气相色谱—质谱法测定水中痕量有机磷和阿特拉津农药[J]. 中国环境监测,2013,29(1):112 - 115. (In English: Wang YL, Zhou Y. Determination of organic phosphorus and atrazine in water by SPME-GC-MS [J]. Environ Monit China, 2013, 29(1): 112 - 115.)

[5] 许秀艳,曹方方,丁曦宁,等. 顶空固相微萃取—气相色谱法测定水中 16 种有机氯农药的含量[J]. 理化检验—化学分册,2017,53(10):1171 - 1176. (In English: Xu XY, Cao FF, Ding XN, et al. GC Determination of 16 organochlorine pesticides in water with headspace solid phase microextraction [J]. Phys Test Chem Anal (Part B: Chem Anal), 2017, 53(10): 1171 - 1176.)

[6] 刘静,曾兴宇,烟卫. 自动顶空固相微萃取气相色谱法同步分析水中 17 种有机物[J]. 分析试验室,2010,29(12):55 - 58. (In English: Liu J, Zeng XY, Yan W. Simultaneous quantization of seventeen organic compounds in water by automated HS-SPME and gas chromatography [J]. Chin J Anal Lab, 2010, 29(12): 55 - 58.)

[7] 王凌,徐晓琴,李庆玲,等. 固相微萃取—气相色谱/质谱(SPME-GC/MS)联用分析海水中痕量有机磷农药[J]. 环境化学,2006,25(1):110 - 114.

[8] 傅若农. 固相微萃取(SPME)近几年的发展[J]. 分析试验室,2015,34(5):602 - 620. (In English: Fu R N. Development of solidphase microextraction (SPME) in recent years [J]. Chin J Anal Lab, 2015, 34(5): 602 - 620.)

[9] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会. GB/T 6682 - 2008 分析实验室用水规格和试验方法[S]. 北京:中国标准出版社,2008.

[10] 朱涛,朱莉萍,周宏琛,等. 固相微萃取—气相色谱—质谱法检测 30 种农药残留[J]. 分析试验室,2013,32(3):115 - 120. (In English: Zhu T, Zhu LP, Zhou HC, et al. Determination of different kinds of pesticides by SPME-GC/MS [J]. Chin J Anal Lab, 2013, 32(3): 115 - 120.)

(责任编辑:陈钰)

【读者·作者·编者】

《环境卫生学杂志》被中国科技核心期刊收录

由国家卫生和计划生育委员会主管,中国疾病预防控制中心主办,中国疾病预防控制中心环境与健康相关产品安全所承办的《环境卫生学杂志》,于 2017 年被中国科技核心期刊(中国科技论文统计源期刊)收录。

《环境卫生学杂志》作为综合性卫生类研究学术期刊,一直致力于建立国家级环境卫生学术交流和信息共享的平台,便于科技人员及时了解本领域的前沿动态和最新进展,拓宽业务人员的视角和思路,为卫生行政部门决策提供科学依据和建议,传播环境卫生信息,积极推进环境卫生事业发展。自创刊以来,始终严把审稿关,对于提升学术质量毫不懈怠。

近年来,在编委会、审稿专家和编辑部的共同努力下,在主办单位、读者、作者的共同支持下,《环境卫生学杂志》的学术水平不断提高。我们正以不断提高的学术水平和编辑出版质量为读者与作者们服务,今后我们将更加的开拓进取,为环境卫生学科的发展尽最大的努力!

(本刊编辑部)