基于时间序列分析的奥运会奖牌数预测*

李　响
（辽宁师范大学附属中学　大连　116023）

摘　要　从奥运会历史数据中发现时序关联，建立预测模型具有深远意义。研究中使用时间序列分析，以法国历届奥运会奖牌数作为实验数据，在R语言下实现了时间序列模型的建立、检验并预测出法国下届奥赛的奖牌数。研究结果表明：将时间序列分析应用于奥运会成绩的挖掘具有一定的科学性，但时间序列分析适用于简单、稳定、具有周期性的数据，为了更充分高效地对奥赛成绩进行预测，应该全面考虑相关因素，并将多种分析方法结合使用。

关键词　时间序列；奥运会；预测；数据挖掘
中图分类号　G812.6　　**DOI**：10. 3969/j. issn. 1672-9722. 2018. 03. 024

Olympic Games Medals Prediction Based on Time Series Analysis

LI Xiang
（The School Affiliated to Liaoning Normal University, Dalian　116023）

Abstract　It is of great significance to establish the prediction model from the historical data of Olympic Games. In the research, the time series analysis is used and the experiment is based on the France's historical data of medals, according to the time series analysis principal, the establish, test, and the prediction of France's medals on the next Olympic Games is realized in R. The results shows: using time series analysis on mining the Olympic Games historical data is somewhat meaningful, but time series analysis is adapted to data that are simple, stable and periodic. To predict more efficiently, it's important to consider all the related factors, and use multiple ways to predict.

Key Words　time series, olympic games, prediction, data mining
Class Number　G812.6

1　引言

奥运会是目前世界上影响力最大的体育盛会，从系统科学的观点看，奥运会赛事是一个动态复杂的大系统，如何科学而准确地建立比赛成绩的预测模型，揭秘奖牌背后的规律，具有广泛而深远的现实意义。随着数据挖掘技术的发展，越来越多的研究者们将不同的算法和工具应用到奥运会奖牌榜的分析、预测中，所建立的预测模型主要分为三类：时间序列预测模型、经验模型及智能化预测模型。如黄昌美、李坤等建立了灰色关联模型对田径、男篮等赛事进行分析和预测[1-2]；张龙、孟刚等进行了奥运会田径项目金牌时空动态演变分析[3]；王国凡、赵武等提出了遗传算法与回归分析相结合的奥运会成绩预测方法[4]；龚剑等研究了基于人工神经网络的奥运会中国男篮成绩预测模型[5]。上述研究为本研究提供了借鉴和参考，但是这些研究所使用的实验环境大都是Matlab、SPSS之类的传统工具，虽然这些工具在统计分析、绘图展示等方面具有强大的功能，但是这些工具都不是开源免费的，使得很多功能的使用受到了限制。为此，本研究选取免费开源且具有丰富算法包的R语言作为实验环境，使用时间序列分析方法挖掘出历届奥运会的历史成绩间的时序关联关系，从而建立预测模型，预测后续奥运会赛事的奖牌数。

2　时间序列分析基本原理

2.1　什么是时间序列及时间序列分析

时间序列是按时间顺序的一组数字序列，时间序列分析就是利用这组数列，应用定量的数理统计

* **收稿日期**：2017年9月8日，**修回日期**：2017年10月26日
作者简介：李响，男，研究方向：C++编程，R语言编程，数据分析。

方法加以处理,以预测未来事物的发展[6]。它的基本原理是:承认事物发展的延续性,应用历史数据可以推测事物的发展趋势;考虑到事物发展的随机性,任何事物发展都可能受到偶然因素影响。

2.2 时间序列分析的关键步骤

2.2.1 数据平稳性检验

时间序列分析的基础是选择合适的数据,通常要求数据是平稳的(或差分后平稳),可以说平稳是时间序列分析非常重要的假设,只有基于平稳时间序列的预测才是有效的,因此平稳性检验是时间序列分析的关键环节。

所谓时间序列平稳指的是:假定某个时间序列由某一随机过程生成,即假定时间序列$\{Xt\}$(t=1,2,…)的每一个数值都是从一个概率分布中随机得到的。如果经由该随机过程所生成的时间序列满足下列条件:

1)均值$E(Xt)=m$是与时间t无关的常数;

2)方差$\mathrm{Var}(Xt)=s^2$是与时间t无关的常数;

3)协方差$\mathrm{Cov}(Xt,Xt+k)=gk$是只与时期间隔k有关,与时间t无关的常数;

则称经由该随机过程而生成的时间序列是(弱)平稳的。该随机过程便是一个平稳的随机过程。

平稳性检验即单位根检验,就是检验序列中是否存在单位根,如果不存在单位根则认为序列是平稳的,检验方法主要包括adf检验、kpss检验、pp检验等[7],adf检验是最常用的方法。adf检验的原假设是存在单位根(即序列不平稳),检验结果如果p值小于0.05则拒绝原假设,认为序列平稳;如果p值大于0.05则接受原假设,认为序列不平稳。

2.2.2 时间序列模型与模型参数的确定

常用的时间序列模型包括ar、ma、arma等,这些模型全部建立在时序平稳的基础上[8]。arma模型的全称为Auto-Regressive and Moving Average Model,即自回归滑动平均模型,它由自回归模型(即ar模型)与滑动平均模型(即ma模型)为基础混合构成。实际应用中考虑到原始数据序列未必稳定,需要进行差分处理,因此引入了改进的arma模型—arima模型。arima模型的具体形式为arima(p,d,q),这里的d是对原时序进行逐期差分的阶数,差分的目的是为了让某些非平稳(具有一定趋势的)序列变换为平稳的,通常来说d的取值一般为0,1,2。当d、q为0时,arima(p,d,q)等价于ar(p)模型;当p,d为0时,arima(p,d,q)等价于ma(q)模型,当d为0时,arima(p,d,q)等价于arma(p,q)模型。参数p、q的值通常通过自相关图(简称acf图)和偏自相关图(简称pacf图)观察判断,具体方法将在本文的第3部分阐述。

3 R语言下基于时间序列的奥运会奖牌预测

R语言是用于统计分析、绘图的开源数据分析软件,由一个庞大且活跃的全球性研究型社区维护。与其它流行的统计软件(如Excel、Matlab、SAS、SPSS)相比,R语言的优势主要体现在:开源免费、易于扩展、数据包丰富、可视化功能强大、可运行于多种平台。

3.1 实验数据的选取

本研究所使用的数据来自国际奥组委官方网站(www.olympic.org),研究中对中国、美国、俄罗斯、英国、法国、德国、意大利等奥运强国的比赛成绩进行了初步分析,认为法国的奥运会奖牌数据最适合用来做时间序列预测研究。因为从1948~2016年,法国每四年都会参加夏季奥运会,从未间断过,历史数据最丰富且具有明显的周期性,如表1所示。

表1 法国历届夏奥会奖牌数

games	gold	silver	bronze
1948	10	6	13
1952	6	6	6
1956	4	4	6
1960	0	2	3
1964	1	8	6
1968	7	3	5
1972	2	4	7
1976	2	3	4
1980	6	5	3
1984	5	7	17
1988	6	6	7
1992	8	6	18
1996	15	7	15
2000	13	14	11
2004	11	9	13
2008	7	18	18
2012	11	11	13
2016	10	18	14

我们将表1的数据保存到d:\france.csv文件中作为实验数据。

3.2 时间序列预测模型的建立

3.2.1 平稳性检验

根据前文所述的时间序列分析的基本原理,首先需要从定量的角度对实验数据进行平稳性检

验[9]。在R语言中可以使用tseries包提供的adf.test()、kpss.test()、pp.test()等函数进行平稳性检验，通常选取一种检验方法即可。平稳性检验的R语句如下：

1）读入数据

```
mydata<-read.csv("d:\\france.csv",header=T)
```

2）选取目标列生成时间序列

表1中的gold、silver、bronze数据列分别代表金牌、银牌、铜牌数，可以任选一列生成时间序列。以silver这一列为例，使用ts()函数将silver列的数据生成时间序列：

```
ps<-ts(mydata$silver)
```

3）绘制时间序列图，从形状上大体判断是否平稳

使用下列语句可以绘制上述时间序列图：

```
plot(ps,main="silver")
```

绘制出的时间序列图如图1所示。从图1可以看出，银牌数据近几届有明显的上升趋势，但也存在起伏，初步断定原始数据并不平稳。

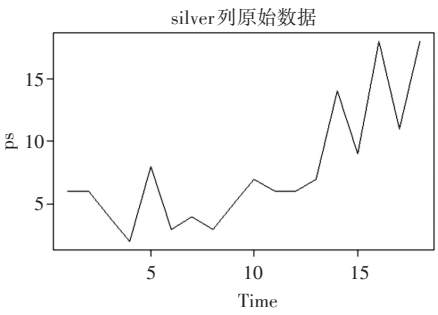

图1　法国历届夏奥会银牌数时间序列

4）使用adf.test()函数检验平稳性

使用adf.test()可以进一步从定量的角度明确数据序列是否平稳。语句如下：

```
library("tseries")
adf.test(ps)
```

该命令的执行结果为

```
Augmented Dickey-Fuller Test
data: ps
Dickey-Fuller = -1.558, Lag order = 2, p-value = 0.7407
alternative hypothesis: stationary
```

对于adf检验，我们可以通过判断结果中的p-value值来确定序列是否平稳，如果p-value小于临界值0.05则认为序列是平稳的。因此，silver列的原始数据不平稳。

如果序列经检验后不平稳，则需要进行差分，直到某阶差分平稳为止[10]；如果最高阶差分后仍不平稳，则认为数据无规律，时间序列分析中止。因此需要对silver列的数据进行差分处理，先做1阶差分，语句如下：

```
d1<-diff(ps,1)
```

绘制差分后的时序图：

```
plot(d1,main="一阶差分")
```

运行结果如图2所示。

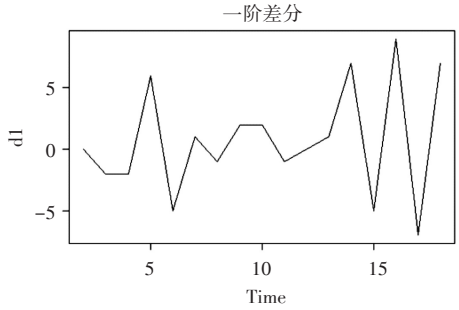

图2　银牌数据一阶差分后的结果

对比图1和图2发现，差分后的数据明显比原始数据平稳了很多。进一步用adf.test()检验差分后的数据是否平稳：

```
adf.test(d1)
```

运行结果为

```
Augmented Dickey-Fuller Test
data: d1
Dickey-Fuller = -3.6109, Lag order = 2, p-value = 0.04922
alternative hypothesis: stationary
```

从结果可以看出p-value小于临界值0.05，可以认为1阶差分后的数据是平稳的，差分的阶数就是arima(p,d,q)模型中参数d的值，因此，可以断定预测模型应该是arima(p,1,q)，下一步的任务是确定p和q的值。

3.2.2　根据自相关图和偏自相关图定阶p、q

arima模型中参数p、q的确定其实是比较复杂的，在实际应用中通常使用观察法，也就是绘制自相关图(acf图)和偏自相关图(pacf图)，如果acf图在q+1处突然截断，则在q处截尾，可确定参数q；同理，如果pacf图在p处截尾，则可确定参数p。

在R语言中可以使用forecast包或stats包中的acf()和pacf()函数来绘制自相关图和偏自相关图，对于上述一阶差分后的平稳数据，使用如下语句：

```
acf(d1)
```

绘制出的自相关图如图3所示。从图3可以看出，自相关图在1阶处超过临界值，2阶之后值逐渐减小，因此，认为q取1比较合适。

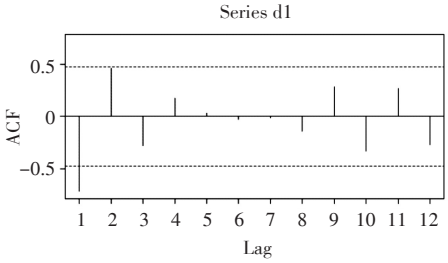

图3 自相关图

接着使用pacf()函数绘制偏自相关图:

pacf(d1)

绘制出的偏自相关图如图4所示。同理,观察偏自相关图认为p取1比较合适。因此确定预测模型为arima(1,1,1)。

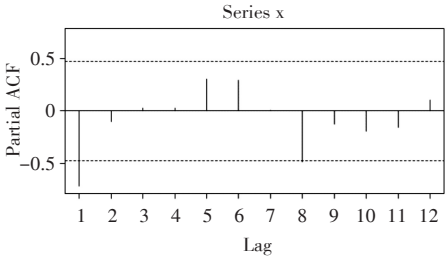

图4 偏自相关图

3.2.3 预测模型的构建及检验

确定好arima模型中的参数后,在R语言中可以使用stats包中的arima()函数构建预测模型,语句如下:

model<−arima(ps,order=c(1,1,1))

构建完模型后,需要对模型进行检验,只有通过检验,才证明是可靠、有效的模型,才能用来进行后续的预测。实质上是对模型残差序列进行白噪声检验。若残差序列不是白噪声,说明还有一些重要信息没被提取,应重新设定模型[11]。通常对残差序列进行白噪声检验使用Ljung−Box检验,在R语言中可以使用stats包中的Box.test()函数进行该项检验,语句如下:

Box.test(model$residuals)

model$residuals表示模型的残差序列,检验结果如下:

Box−Pierce test

data: model$residuals

X−squared = 0.2588, df = 1, p−value = 0.6109

从结果可以看出,p−value大于临界值0.05,所以认为模型的残差序列为白噪声序列,模型通过检验,建立成功。

3.2.4 使用模型进行预测

模型建立成功后,可以用来进行预测,比如预测下一次奥运会法国的银牌数目,在R语言下可使用如下语句:

predict(model,n.ahead=1)

预测结果如下:

$pred

Time Series:

Start = 19

End = 19

Frequency = 1

[1] 12.48319

$se

Time Series:

Start = 19

End = 19

Frequency = 1

[1] 2.95902

上面结果中,变量$pred表示预测值,变量$se为误差。

4 结语

本研究使用时间序列分析挖掘奥运会历史成绩中存在的时序关联关系,并以法国历届夏奥会的银牌数为实验数据,在R语言下根据时间序列分析原理,建立了预测模型并作了模型检验,成功预测出下一届赛事法国的银牌数。使用同样的方法也可以对金牌数和铜牌数进行预测。研究结果表明:

1)对于简单、稳定或周期性的数据,使用时间序列分析建立预测模型具有较好的效果。但是,时间序列分析并不适用于任何数据,比如,对于有明显上升趋势的中国体育竞赛成绩而言,使用时间序列预测,预测值可能会低于实际值。

2)奥运会比赛成绩的影响因素很多,除了可以从历史数据中找寻规律外,还应该全面考虑综合国力、东道主效应等其他因素,使用神经网络、遗传算法等智能方法进行更加完善的预测。这也是本研究的后续研究方向。

参考文献

[1] 黄昌美. 奥运会田径运动成绩的灰预测建模及其变化发展趋势分析[D]. 湘潭:湖南科技大学,2012.
HUANG ChangMei. Grey Prediction modeling and Development Trend Analysis of Olympic Track and Field Events Achievement[D]. Xiangtan: Hunan University of Science and Technology, 2012.

(下转第545页)

CHEN Jianhong, CHEN Kefei, LONG Yu, et al.Ciphertext Policy Attributes-Based Parallel Key-Insulated Encrytion [J]. Journal of Software, 2012, 23(10): 83-92.

[8] Foster I. Globus Toolkit Version 4: Software for Service-Oriented Systems [J]. Journal of Computer Science and Technology, 2006, 21(4): 513-520.

[9] 刘萍萍, 闫琳英. 面向云存储的访问控制系统的研究[J]. 西安工业大学学报, 2015, 35(5): 355-359.
LIU Pingping, YAN Linying. Research on cloud storage oriented access control system [J]. Journal of Xi'an Technological University, 2015, 35(5): 355-359.

[10] 周彦萍, 马艳东. 基于CP-ABE访问控制系统的设计与实现[J]. 计算机科学与发展, 2014, 24(2): 145-148.
ZHOU Yanping, MA Yandong.Design and implementation of access control system based on CP-ABE[J]. Computer science and development, 2014, 24(2): 145-148.

[11] M.Pirretti, P.Traynor, P.McDaniel, and B.Waters.Secure Attribute-Based Systems.

[12] 孙旭. 基于CP-ABE算法的云数据服务访问控制系统的设计与实现[D]. 成都: 电子科技大学, 2015.
SUN Xu.Design and implementation of cloud data service access control system based on CP-ABE algorithm [D]. Chengdu: University of Electronic Science and technology, 2015.

[13] 程相然, 陈性元, 张斌, 等. 基于属性的访问控制策略模型[J]. 计算机工程, 2010, 36(15): 131-133.
CHEN Xiangran, CHEN Xingyuan, ZHANG Bin, et al.Attribute-Based Access Control Policy Model[J]. Computer Engineering, 2010, 36(15): 131-133.

[14] P Sasikala.Cloud computing: present status and future implications [J]. International journal of cloud computing. 2011, 1(1): 23-26.

[15] 王鹏. 走近云计算[M]. 北京: 人民邮电出版社, 2009.
WANG Peng. Cloud Computing[M]. Beijing: Post & Telecom Press, 2009.

(上接第 536 页)

[2] 李坤. 第29届奥运会中国男篮技术指标的灰色关联分析[D]. 北京: 北京体育大学, 2010.
LI Kun. Grey Incidence Analysis of Technical Statistic of Chinese Men's Basketball Team in the 29th Olympic Games[D]. Beijing: Beijing Sport University, 2010.

[3] 张龙, 孟刚, 郭朝廷. 奥运会田径项目金牌时空动态演变分析[J]. 中国体育科技, 2013, 49(5): 17-27.
ZHANG Long, MENG Gang, GUO Chanting. Dynamic Evolution of Gold Medal Time and Space of Olympic Games Athletics[J]. China Sport Science and Technology, 2013, 49(5): 17-27.

[4] 王国凡, 赵武, 刘徐军, 等. 基于GA和回归分析的奥运会成绩预测研究[J]. 中国体育科技, 2011, 47(1): 4-8, 16.
WANG Guofan, ZHAO Wu, LIU Xujun, et al. Olympic Performance Prediction based on GA and Regression Analysis [J]. China Sport Science and Technology, 2011, 47(1): 4-8, 16.

[5] 龚剑. 基于人工神经网络2008奥运会中国男篮成绩预测实验研究[D]. 武汉: 武汉体育学院, 2007.
GONG Jian. An Empirical Analysis of The Chinese Men's Basketball Achievement Prediction in The Olympic Games of 2008 Based on ANN[D]. Wuhan: Wuhan Institute of Physical Education, 2007.

[6] 首招勇, 杨媛媛. 时间序列问题的建模方法和过程[J]. 数学理论与应用, 2012, 32(1): 112-120.
SHOU Zhaoyong, YANG Yuanyuan. On the Modelling of Time Series [J]. Mathematical Theory and Applications, 2012, 32(1): 112-120.

[7] 管河山, 邹清明, 罗智超. 时间序列平稳性分类识别研究[J]. 统计与信息论坛, 2016, 31(4): 3-8.
GUAN Heshan, ZOU Qingming, LUO Zhichao. Study on Classification and Identification of Time Series Stationarity [J].Statistics & Information Forum, 2016, 31(4): 3-8.

[8] 王娜. 时间序列建模、预报的原理[J]. 吉林工程技术师范学院学报, 2012, 23(3): 78-80.
WANG Na. Time Series Modeling, Forecast Principle [J]. Journal of Jilin Teachers Institute of Engineering and Technology, 2012, 23(3): 78-80.

[9] 刘罗曼. 时间序列平稳性检验[J]. 沈阳师范大学学报(自然科学版), 2010, 28(3): 357-359.
LIU Luoman. Checking of Time Series Stationarity [J]. Journal of Shenyang Normal University (Natural Science), 2010, 28(3): 357-359.

[10] 麦鸿坤, 肖坚红, 吴熙辰, 等. 基于R语言的负荷预测ARIMA模型并行化研究[J]. 电网技术, 2015, 39(11): 3216-3220.
MAI Hongkun, XIAO Jianhong, WU Xichen, et al. Research on ARIMA Model Parallelization in Load Prediction Based on R Language[J]. Power System Technology, 2015, 39(11): 3216-3220.

[11] 刘瑶. 基于ARMA模型的人民币汇率预测研究——以人民币兑美元汇率为例[J]. 廊坊师范学院学报(自然科学版), 2016, 16(2): 53-58.
LIU Yao. Research on RMB Exchange Rate Prediction Based on ARMA Model——A Case Study of RMB against USD [J]. Journal of Lang fang Teachers University (Natural Science Edition), 2016, 16(2): 53-58.