基于缺失值的海南旅游总收入的季节ARIMA模型及预测

陈传钟，汪文俊*，缪光美
(海南师范大学 数学与统计学院,海南 海口 571158)

摘 要:讨论了带有缺失值的2007年1月至2013年2月的海南旅游总收入的数据,利用不同处理缺失值的方法对数据进行整合,得到海南省旅游总收入服从季节ARIMA模型,并由此对海南旅游总收入趋势进行有效预测.

关键词:缺失值;时间序列;ARIMA模型
中图分类号:O 213 **文献标识码**:A **文章编号**:1674-4942(2014)01-0001-06

Seasonal ARIMA Model and Predict on the Total Income of Tourism in Hainan with Missing Values

CHEN Chuanzhong, WANG Wenjun*, MIAO Guangmei
(*College of Mathematics and Statistics, Hainan Normal University, Haikou* 571158, *China*)

Abstract: This paper discusses the tourism revenue data of Hainan with missing values from January 2007 to February 2013. We combine data by using different methods for dealing with missing values, and find that the Seasonal ARIMA model could describe variation tendency of the tourism revenue of Hainan.

Key words: Missing values; Time series; ARIMA model

随着旅游业在世界各地的迅猛发展,有关旅游业可持续发展的研究越来越受到人们的重视,近年来,定量分析及统计方法被广泛应用到旅游发展研究中,本文拟采用时间序列的方法对含缺失值的海南旅游总收入数据,统计建模,并对海南旅游收入的具体情况进行分析讨论.

时间序列分析是一种对动态数据处理的时域参数方法,目的是研究所给的动态数据序列的统计规律,以用于解决实际问题. 海南旅游收入变化趋势受到季节影响,每年7月-9月,10-12月、1月都是高峰期,然而并没有文献对变化的趋势具体研究,本文考虑利用"海南省旅游政务网"[1]提供的可靠数据,对海南省旅游总收入进行处理和预测.

1 缺失值处理

由于2011年1月和2011年2月数据缺失,首先考虑序列均值、临近点的均值、临近点的中位数、线性插值法、点处的线性趋势五种不同方法对缺失值进行处理. 通过先期的计算比较,最终选定临近点的中位数、线性插值法两种方法[2].

临近点的中位数表示缺失值邻近的几个点的中位数,具体几个点由附近点的跨度来决定. 临近点中值弥补缺失值前后对比见图1、图2.

线性插值法表示应用线性插值法填补缺失值,即缺失值前一个数据和后一个数据建立插值直线,然后找到缺失点在线性插值函数的函数值作为该缺失值,线性插值法弥补缺失值前后对比见图3、图4.

从以上对比图可以看到,临近点的中位数插值法和线性插值法都能很好的拟合原始数据的变化趋势. 下节我们将利用此两种方法获得的完整数据进行统计建模.

收稿日期:2013-12-04
基金项目:海南省自然科学基金项目(110003)
*通讯作者

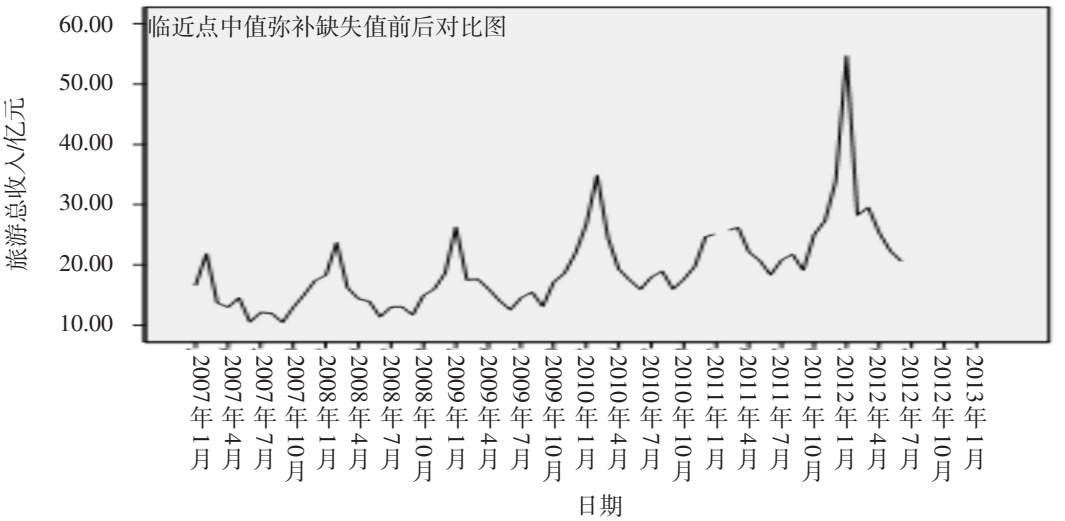

图1 弥补缺失值前的序列图

Fig.1 Graph of time series without imputation for missing values

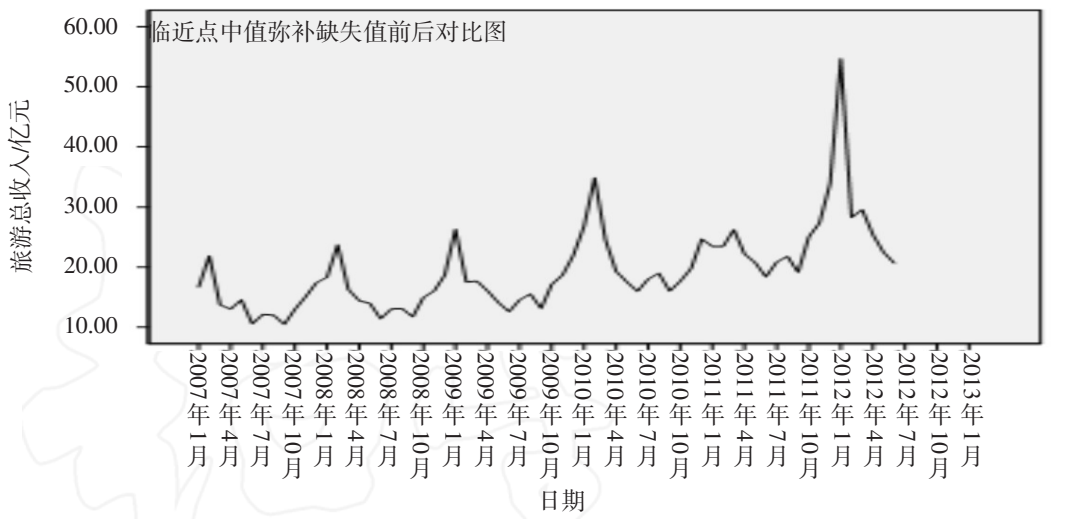

图2 弥补缺失值后的序列图

Fig.2 Graph of time series with imputation for missing values

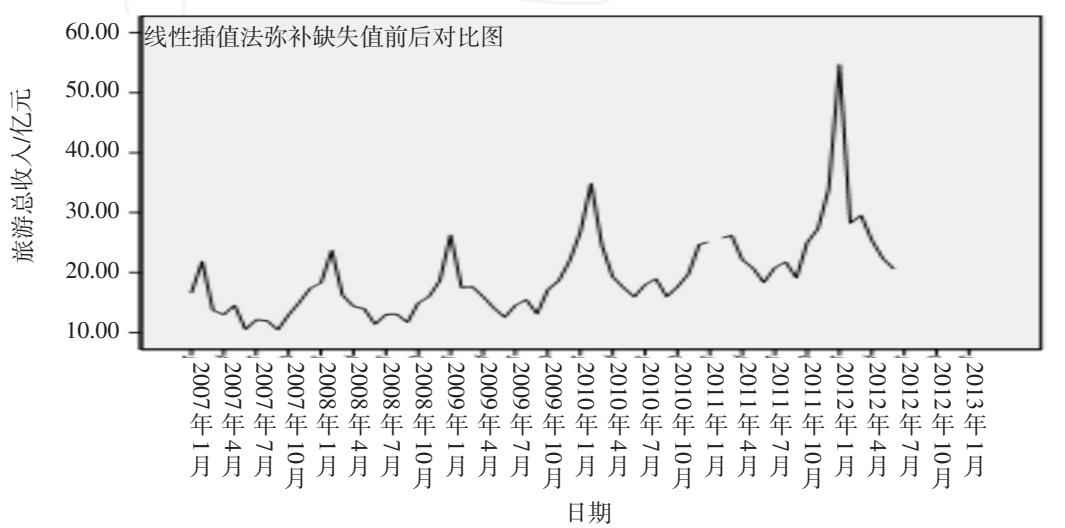

图3 弥补前的序列图

Fig.3 Graph of time series without imputation for missing values

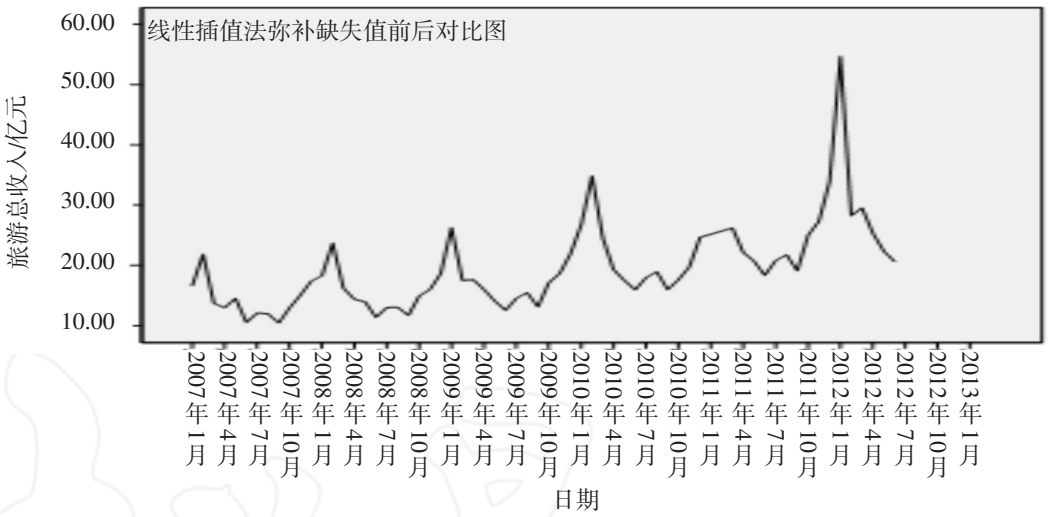

图4　弥补后的序列图

Fig.4　Graph of time series with imputation for missing values

2　数据统计建模

2.1　模型建立的理论基础

从图2和图4中观察到数据具有明显的周期性(以12个月为一周期),和趋势性,因此可以尝试时间序列的季节ARIMA$(p,d,q)(P,D,Q)^s$(可乘季节ARIMA)[3-4]进行拟合.

一个一般的具有非平稳(通常的)阶数p,d,q,季节阶数P,D,Q及周期s可乘季节ARIMA(SARIMA)模型为

$$\phi_p(L)\Phi_p(L)W_t=\theta_q(L)\Theta_Q(L)Z_t,$$

具体结构如下:

$$\begin{cases}\phi_p(x)=1-\alpha_1x-\alpha_2x^2-\cdots-\alpha_px^p\\ \Phi_p(x)=1-C_1x^s-C_2x^{2s}-\cdots-C_px^{ps}\\ \theta_q(x)=1+\beta_1x+\beta_2x^2+\cdots+\beta_px^p\\ \Theta_Q(x)=1+T_1x^s+T_2x^{2s}+\cdots+T_Qx^{Qs}\end{cases}$$

差分序列为$W_t=\nabla^d\Delta_s^DX_t$.

2.2　对临近点中值处理的数据进行建模

1、根据图2的趋势性和周期性,对数据做一次季节性差分和一阶逐期差分,观察自相关图和偏自相关图,确定ARIMA模型的相关系数(见图5、图6).

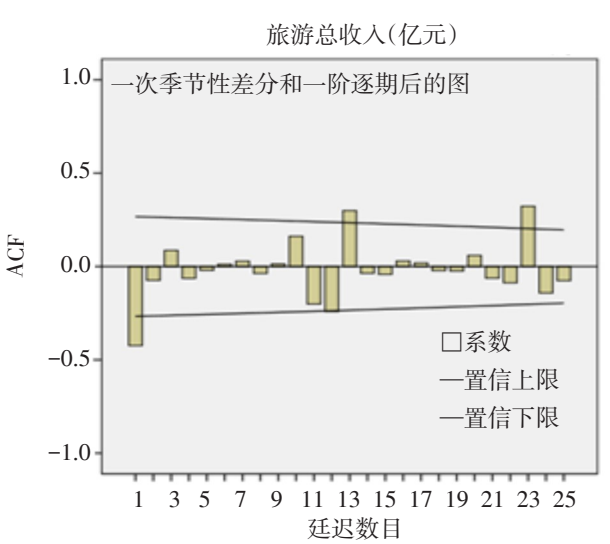

图5　自相关图

Fig.5　ACP

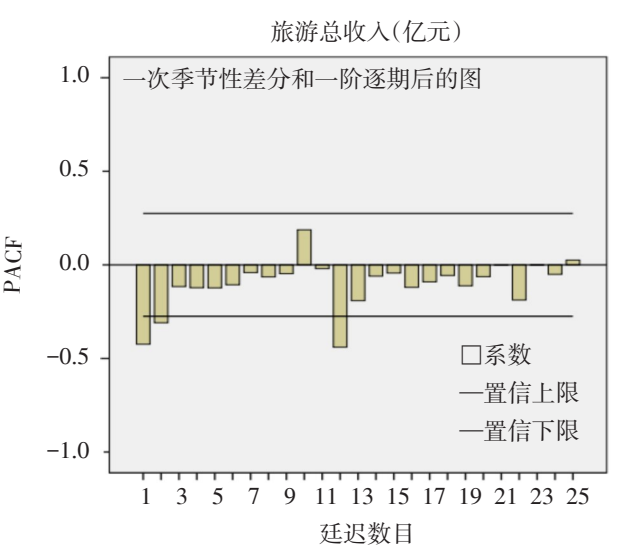

图6　偏自相关图

Fig.6　PACP

从自相关图(ACP)中,看到自第二个延迟数目开始,自相关落在虚线内,一阶以后函数值明显趋于0,呈拖尾性,因此取q=2. 同时,第13阶显著不为0,因此取Q=2.

偏自相关图中,前两阶函数显著不为0,之后趋于0并呈拖尾性,因此取p=3,而第12阶显著不为0,取P=1.

因为以上讨论的是一阶季节性差分和一阶逐期差分,所以取D=1,d=1. 又从图7中可以看到,序列图稳定,所以可以构建模型ARIMA(3,1,2)(1,1,2),S是季节周期,它的取值为4式12.

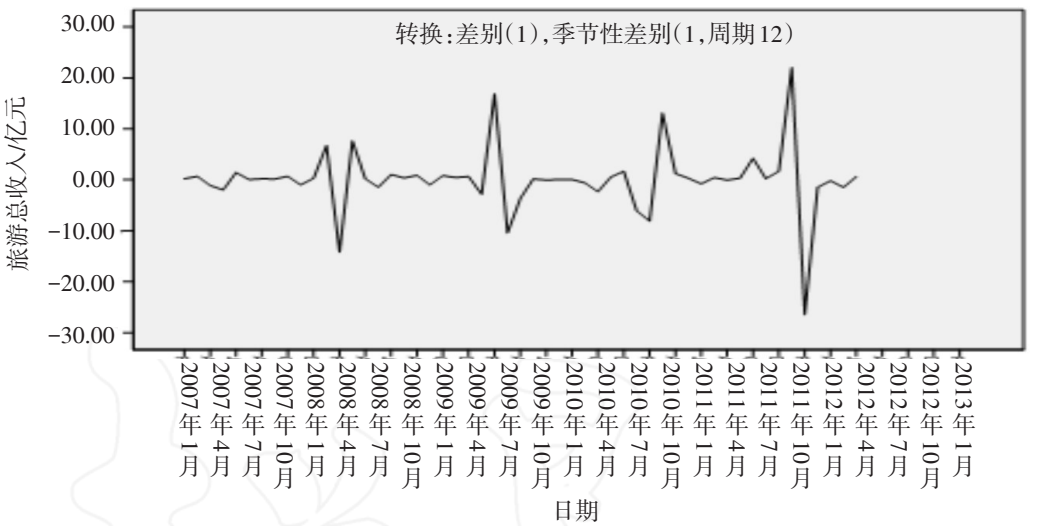

图7 一次逐期差分和一次季节性差分后的序列图

Fig.7 Graph of time series with first order successive and first order seasonal difference

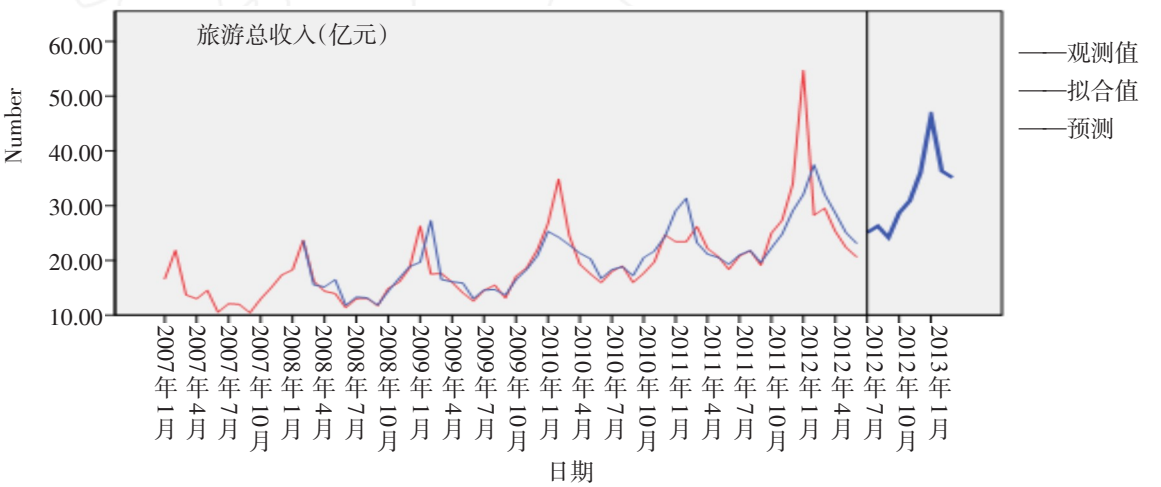

图8 模型拟合图

Fig.8 Model fitting diagram

2)按照所求参数进行建模,得到图形见图8.

明显看到ARIMA(3,1,2)(1,1,2)拟合的效果尚佳.

2.3 对线性插值法处理的数据进行建模

线性差值法的数据处理步骤同上,相关图形数据见图9、图10、图11.

以上的图表中得到的线性插值法的模型为ARIMA(3,1,2)(1,1,2).

3 比较分析及预测

3.1 两种方法模型的比较

从表1、表2,观察到,线性插值法的平稳R方0.651>0.519(临近点中值法),p值0.582>临近点中值的0.286,而正态化的BIC模型值小于临近点中

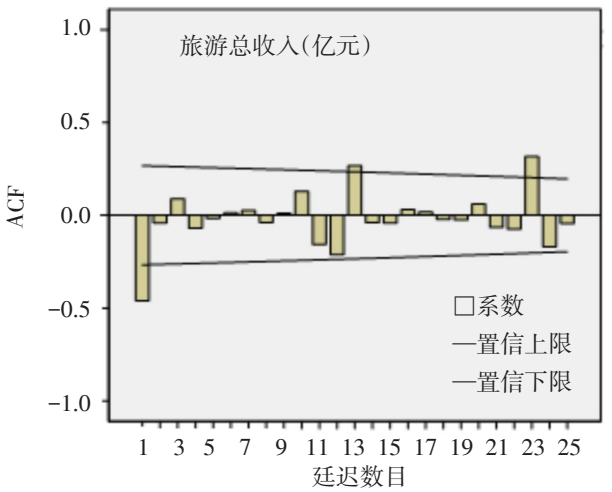

图9　自相关图

Fig.9　ACP

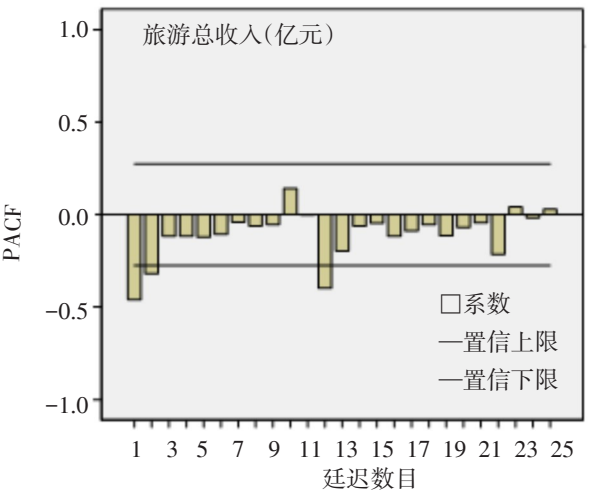

图10　偏自相关图

Fig.10　PACP

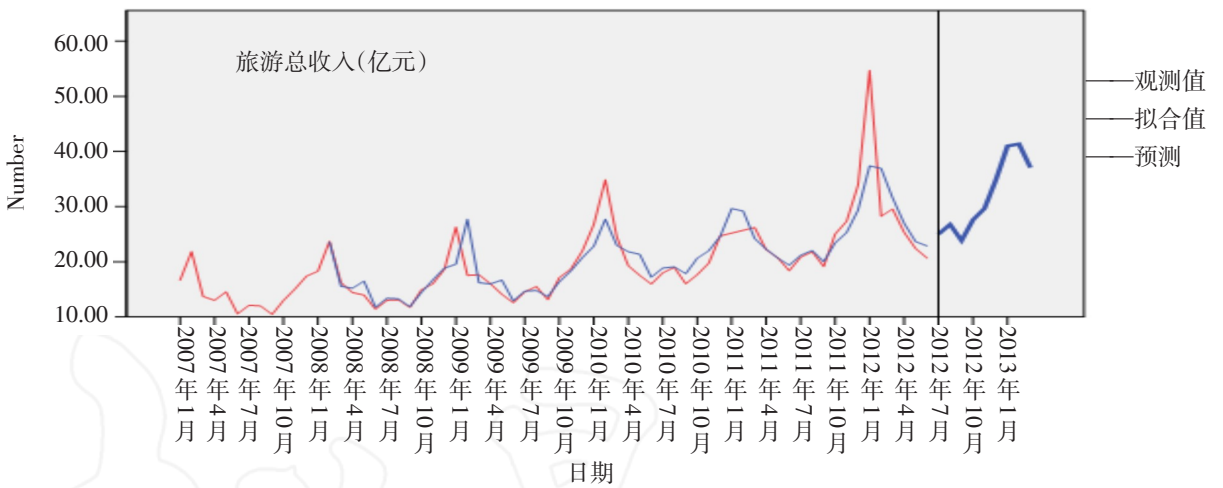

图11　模型拟合图

Fig.11　Model fitting diagram

表1　临近点中值的模型统计量

Tab.1　Model statistics with near median value

模型	预测变量数	模型拟合统计量		Ljung-Box Q(18)			离群值数
		平稳的R方	正态化的BIC	统计量	DF	Sig.	
MEDIAN(五、旅游总收入(亿元),2)-模型_1	0	0.519	3.868	11.988	10	0.286	0

表2　线性插值法的模型统计量

Tab.2　Model statistics with linear interpolation

模型	预测变量数	模型拟合统计量		Ljung-Box Q(18)			离群值数
		平稳的R方	正态化的BIC	统计量	DF	Sig.	
MEDIAN(五、旅游总收入(亿元),2)-模型_1	0	0.651	3.588	7.534	9	0.582	0

值,因此判断线性插值法所得的模型更佳,其模型为ARIMA(3,1,2)(1,1,2).

3.2　基于线性插值法处理缺失值的预测结果

从表中可以看到预测较实际值误差较小,但是从2012年10月到2012年12月预测值偏高,根据2011年同期数据的比较,在表5中发现,2012年整体数据上升趋势并没有2011年那么明显,说明目前国内海南游人数出现一定的疲软状况. 因此该模

表3 实际值与预测值的对比

Tab.3 Comparison of actual and predicted values

	线性插值法的预测值							
模型	201207	201208	201209	201210	201211	201212	201301	201302
预测	25.01	26.77	23.84	27.61	29.66	34.92	40.99	41.34
实际值预测	24.41	24.48	23.5	32.18	34.62	40.94	35.34	34.98
值较实际值变化	2.46%	9.34%	1.45%	−14.19%	−14.32%	−14.71%	15.98%	18.18%

表4 海南入境游人数表

Tab.4 The number of inbound in Hainan

	2010年						2012年	
	7月	8月	9月	10月	11月	12月	1月	2月
2010	17.97	18.98	15.98	17.66	19.74	24.67	–	–
2011	20.82	21.78	19.11	25.01	27.32	33.94	54.71	28.24
2012	24.41	24.48	23.50	32.18	34.62	40.94	35.34	34.98

表5 入境游人数对比

Tab.5 Contrast of the number of inbound

	2010年						2012年	
	7月	8月	9月	10月	11月	12月	1月	2月
11年同期较10年	15.86%	14.75%	19.59%	41.62%	38.40%	37.58%	–	–
12年同期较11年	17.24%	12.40%	22.97%	28.67%	26.72%	20.62%	−35.40%	23.87%

表6 文章所用的数据

Tab.6 The data

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
2007	16.65	21.86	13.7	12.99	14.54	10.54	12.11	11.98	10.45	12.96	15.05	17.37
2008	18.29	23.71	16.22	14.39	13.95	11.38	13.01	13.08	11.71	14.91	16	18.63
2009	26.29	17.51	17.64	16	14.09	12.55	14.54	15.5	13.11	17.12	18.67	21.94
2010	26.76	34.86	24.55	19.28	17.52	15.94	17.97	18.98	15.98	17.66	19.74	24.67
2011	–	–	26.21	22.17	20.72	18.36	20.82	21.78	19.11	25.01	27.32	33.94
2012	54.71	28.24	29.55	25.33	22.39	20.65	24.41	24.48	23.5	32.18	34.62	40.94

型按照趋势拟合具有一定的误差,但是在允许的范围内.

4 结论

海南旅游总收入受到季节的影响,本文基于线性插值法处理缺失值的数据,建立的季节ARIMA(3,1,2)(1,1,2)模型,较为准确的拟合海南省旅游总收入的变化趋势,其预测值亦可以为研究海南旅游变化动态提供参考意见.

参考文献:

[1] 海南省旅游发展委员会[EB/OL]. [2013−10−04]http://tourism.hainan.gov.cn/goverment/govPrePic/govBelow-Pic1/.

[2] 薛薇. spss统计分析方法及应用[M]. 2版. 北京:电子工业出版社,2011:454−462.

[3] 王燕. 应用时间序列[M]. 3版. 北京:中国人民大学出版社,147−148.

[4] 吴喜之. 复杂数据统计方法−基于R的应用[M]. 北京:中国人民大学出版社,2012:176−181.

责任编辑:毕和平