DOI:10.13546/j.cnki.tjyjc.2017.06.006

复杂季节时间序列模型研究

马佳羽，韩兆洲
（暨南大学 经济学院，广州 510632）

摘 要：季节时间序列有时不止有一个季节周期，比如以小时计的数据，24小时可以是一个季节周期，同时，一周可以是一个季节周期。为解决传统模型不能处理复杂季节问题，文章采用傅里叶级数序列作为ARIMA模型的辅助回归元，对我国2004年1月至2015年8月的铁路客运量进行拟合。结果表明，分别选择正余弦个数为1和4的2.6和12个月为周期的傅里叶级数作为辅助回归元拟合ARIMA(3,1,1)模型最优，拟合的平均绝对百分比误差（MAPE）为5.46%。在此基础上对我国2016年各月份的客运量进行了预测。

关键词：复杂季节；傅里叶级数；铁路客运量
中图分类号：F201 **文献标识码**：A **文章编号**：1002-6487(2017)06-0027-04

0 引言

季节性时间序列由于其往往不仅具有趋势性，而且具有季节性，因此比较难研究和预测，尤其在大数据时代，时间序列数据更加多样化，大量时间序列不仅可以以年、月、季记录，还可以方便地以日、小时，甚至分钟记录，比如网站浏览量。这些时间序列很多带有季节性，并且不止一个季节周期，比如以小时计的时间序列数据，每12个月是一个季节周期，每一季度内也有季节周期性，甚至一周7天，一天24小时内也有季节性，并且周期数也有可能不是整天或月，比如2.5个月等。研究季节性时间序列的模型如：季节性因子分解预测法、Parsons连环比例法、季节性指数平滑法、Holt-Winters模型、Box-Jenkins季节模型预测法等，但这些模型只能处理带有一种季节周期的情况，而且季节长度一般不超过24个单位，对于类似一周168小时这样的长周期不能合适的处理。对于复杂季节时间序列模型和非整周期数不能很好的处理。本文用傅里叶序列作为ARIMA模型的辅助回归元，建立带有傅里叶序列的ARIMA模型，解决了复杂季节时间序列不好处理的问题，利用AIC准则选择最优模型，做科学预测，并用R语言进行实现。

1 数据来源及特征描述

选用2004年1月至2015年8月的月度铁路运输量数据作为研究对象，数据来源于国家统计局网站的月度数据查询，数据详情见表1。

基金项目：国家社会科学基金资助项目（15ATJ001）
作者简介：马佳羽（1991—），女，青海西宁人，硕士研究生，研究方向：经济预测与决策。
韩兆洲（1955—），男，江苏苏州人，教授，博士生导师，研究方向：经济预测与决策。

Simulation and Testing of Loss Simulation Model

Di Na[1,2], Lu Zhiyi[3]

(1. School of Economics, Tianjin University of Finance and Economics, Tianjin 300222, China;
2. School of Economics & Management, Tianjin Agricultural University, Tianjin 300384, China;
3. School of Science, Tianjin University of Commerce, Tianjin 300134, China)

Abstract: The CAS(Casualty Actuarial Society) has developed a open-source software-LSM(Loss Simulation Model)-which produces claims and transaction data to help the actuaries to select the optimal method of loss reserving with different data, but it has to be tested whether the simulation data is produced in accordance with the specified parameters and distributions. We have tested the distribution of claims number, the trend of severity and the correlation structure of claims number between different lines used different methods of parameters estimation and goodness of fit in this paper. The results showed LSM could produce consistent simulations with the distribution of claims number and the severity trend, but it is not robust in simulation of correlation structure of claims numbers between different lines.

Key words: loss simulation model; claims numbers; severity; testing of goodness of fit;

表1　我国铁路客运量月度数据(2004.1-2015.8)　(单位:亿人)

时间	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
2004	0.78	0.93	0.87	0.8	0.85	0.73	0.91	0.96	0.83	0.86	0.73	0.74
2005	0.930	1.060	0.930	0.910	0.970	0.860	1.080	1.120	0.940	1.000	0.860	0.850
2006	1.070	1.130	0.990	0.990	1.070	0.960	1.200	1.220	1.020	1.100	0.927	0.920
2007	0.990	1.109	1.197	1.026	1.140	1.020	1.310	1.350	1.145	1.208	1.030	1.070
2008	1.190	1.285	1.186	1.162	1.166	1.147	1.379	1.406	1.250	1.257	1.080	1.033
2009	1.328	1.359	1.180	1.249	1.289	1.152	1.419	1.501	1.218	1.358	1.107	1.089
2010	1.272	1.422	1.409	1.327	1.378	1.336	1.600	1.620	1.382	1.528	1.235	1.219
2011	1.520	1.572	1.411	1.555	1.531	1.508	1.816	1.786	1.614	1.626	1.341	1.315
2012	1.647	1.557	1.446	1.645	1.488	1.623	1.798	1.852	1.691	1.509	1.419	1.482
2013	1.876	1.404	1.685	1.750	1.623	1.804	1.993	2.029	1.920	1.641	1.556	1.738
2014	1.905	1.598	1.805	1.984	1.904	1.946	2.239	2.352	2.099	1.792	1.706	2.243
2015	1.785	1.929	2.155	2.109	2.122	2.061	2.478	2.554	-	-	-	-

将表1中的数据画成折线图以便于分析季节周期，如图1所示。

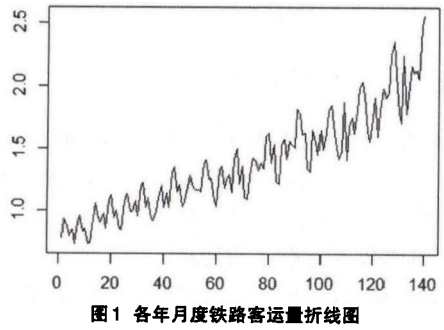

图1 各年月度铁路客运量折线图

从2007年1月到2015年8月铁路客运量时序图可以看出时序既有递增趋势性，又有季节性，其中一个明显的季节周期为一年12个月。分别将每年12个月的客运量绘制在一张折线图中分析，见图2：

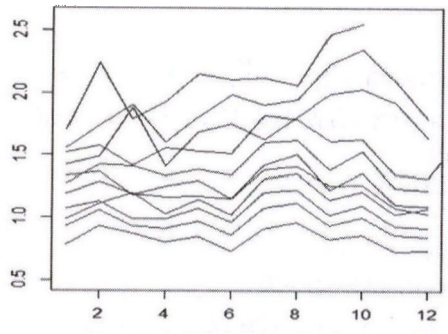

图2 分年月度铁路客运量折线图

图2是将2007—2015年每个月的客流量绘制在一张图中可以看到2004—2006年1-3月中的2月，2008—2014年的1月和2007年与2015年的3月由于正值春节，大批在外工作人员返乡过节，客流量达到最高；由于2008年前五一劳动节放假7天，有很多游客外出旅游，探亲，所以前些年5月份会迎来一次客流高峰，2008年后4月份的清明节放假三天，五一放假三天，因此近几年4月份的旅客较多，因为清明节有独特的传统节日意义，很多在外人员需要回乡，近些年5月份反而是4-6月中的低谷，6月和4月份的旅客数量差不多；7-9月中由于8月接近开学且7-8月正值暑假，大批在外上学的学生需要返校且一定规模的游客会在暑假外出旅行，8月份的游客数量达到最高，且7-8月份的客流量通常会达到全年最高；10-12月由于十一国庆放假7天，10月份客流量最高。通过上文分析，12月肯定是一个长周期，每年12个月中会有2个月或3个月的短季节周期性，由于一个月的客流量大也许是由于有些天的客流量大，而不是月内每天均匀，所以周期本身不需要很精确，因此本文将以0.1个月作为步长，以2.1~3个月作为周期范围，将在下文的模型分析中选择最优的短周期，得到12个月内的一个季节周期。

2　复杂季节时间序列模型

复杂季节ARIMA模型的原理如下：

首先，傅里叶级数的公式为：

$$y_t=\sum_{i=1}^{M}\sum_{k=1}^{k_1}\left[\alpha_{ik}\sin\left(\frac{2\pi kt}{pi}\right)+\beta_{ik}\cos\left(\frac{2\pi kt}{pi}\right)\right] \quad (1)$$

其中p表示季节周期数，k表示正余弦的个数，M表示周期个数。

令ARIMA(p,d,q)模型为：

$$\nabla^d Y_t=\theta_0+\sum_{i=1}^{p}\phi_i\nabla^d Y_{t-i}+\varepsilon_t+\sum_{j=1}^{q}\theta_j\varepsilon_{t-j}$$

即：

$$N_t=Y_t=\theta_0+(1+\phi_1)Y_{t-1}+(\phi_2-\phi_1)Y_{t-2}+\cdots+(\phi_p-\phi_{p-1})$$

$$Y_{t-p}-\phi_{t-p-1}+\varepsilon_t+\sum_{j=1}^{q}\theta_j\varepsilon_{t-j} \quad (2)$$

其中d表示差分阶数，p表示自回归阶数，q表示移动平均阶数。

则复杂季节时间序列模型为：

$$Y_t=N_t+\sum_{i=1}^{M}\sum_{k=1}^{K_4}[\alpha_{ki}\sin\left(\frac{2\pi kt}{pi}\right)+\beta_{ik}\cos\left(\frac{2\pi ki}{pi}\right)] \quad (3)$$

上式中 pi 表示季节周期，比如一天24小时，一周7天168小时等。k表示正余弦的个数，M表示季节周期的个数。

3　模型的选择与估计

模型中需要确定的参数有：ARIMA模型的自回归阶数p、差分阶数d和移动平均阶数q、短季节周期数s以及两个季节周期的傅立叶级数中正余弦的个数k_1和k_2，这些参数的确定都可以通过反复实验比较AIC来确定，首先编写程序产生每种季节周期的傅里叶级数序列，作为ARIMA模型的辅助回归元，由数据的描述性分析可以发现铁路客运量有两个性季节周期，分别是s(待确定)个月和12个月，那么以s个月为周期和12个月为周期。

然后，是 k_1，k_2 和自回归阶数与移动平均阶数的选择，一般k不超过4，因此本文可以尝试k_1从1到4，k_2从1到4的共计 $C_4^2=16$ 种组合，与s组合共160种情况，将每种组合代入ARIMA复杂季节模型。由于通常不止一组(p,q)适合模型且能够通过检验，故根据文献资料，可直接利用AIC准则选择适合的阶数。在R语言中编写循环试验程序产生相对应的160种傅立叶级数，函数内再直接带入R语言的auto.arima函数，该函数可以按AIC准则直接找到最优模型，最后将160个模型的AIC比较，找出最小的AIC对应的

s和k_1/k_2的组合，由于正余弦函数的周期性，某些模型的外部回归元产生完全多重共线性，导致模型失效，无法估计，故有130个有效模型，各组合的AIC值如下页表2所示：

表2　不同短周期和k1/k2组合的AIC值表

s	k_1/k_2	1	2	3	4	s	k_1/k_2	1	2	3	4
2.1	1	-168.460	-154.449	-152.561	-159.432	2.6	1	-167.711	-163.831	-161.361	-157.635
	2	-221.132	-217.202	-216.648	-214.947		2	-222.483	-218.566	-216.859	-214.153
	3	-221.731	-217.815	-217.321	-215.640		3	-223.015	-219.281	-217.627	-214.995
	4	-226.415	-222.482	-223.058	-223.089		4	-236.633	-225.681	-225.107	-222.323
2.2	1	-165.434	-165.318	-162.988	-153.341	2.7	1	-165.906	-167.052	-163.549	-163.885
	2	-218.865	-223.828	-220.242	-216.998		2	-219.547	-228.183	-224.876	-222.468
	3	-220.031	-223.169	-219.532	-216.086		3	-220.662	-227.249	-223.933	-221.850
	4	-225.990	-229.483	-226.140	-222.630		4	-226.815	-233.529	-230.253	-231.518
2.3	1	-158.438	-154.794	-151.076	-164.028	2.8	1	-166.433	-162.696	-160.549	-159.159
	2	-223.859	-220.424	-216.549	-222.311		2	-220.091	-216.615	-215.574	-214.739
	3	-225.458	-222.071	-218.230	-223.926		3	-221.099	-217.805	-215.747	-215.737
	4	-231.260	-227.851	-224.234	-230.313		4	-227.807	-224.820	-218.201	-222.563
2.4	1	-162.672	-228.455	-229.936	-235.633	2.9	1	-165.346	-161.416	-159.483	-155.632
	2	-228.455	NULL	NULL	NULL		2	-221.846	-217.966	-215.330	-211.150
	3	-229.935	NULL	NULL	NULL		3	-222.334	-218.451	-215.852	-212.068
	4	-229.488	NULL	NULL	NULL		4	-230.639	-226.835	-224.167	-220.557
2.5	1	-167.525	-163.713	NULL	NULL	3.0	1	-178.080	NULL	NULL	NULL
	2	-222.334	-218.873	NULL	NULL		2	-229.417	NULL	NULL	NULL
	3	-222.962	-219.586	NULL	NULL		3	-229.969	NULL	NULL	NULL
	4	-229.034	-225.570	NULL	NULL		4	NULL	NULL	NULL	NULL

函数同时确定了对应的最优的ARIMA差分、自回归和移动平均的阶数，至此，模型的形式已经确定下来了，即ARIMA(3,1,1),s=2.6,k_1=1,k_2=4：

$$Y_t=\theta_0+(1+\phi)Y_{t-1}+(\phi_2-\phi_1)Y_{t-2}+(\phi_3-\phi_2)Y_{t-3}-\phi_3Y_{t-4}+\varepsilon_t-\theta\varepsilon_{t-1}+\alpha_{11}\sin(\frac{2\pi t}{2.6})+\beta_{11}\cos(\frac{2\pi t}{2.6})+\alpha_{21}\sin(\frac{2\pi t}{12})+\beta_{21}\cos(\frac{2\pi t}{12})+\alpha_{22}\sin(\frac{4\pi t}{12})+\beta_{22}\cos(\frac{4\pi t}{12})+\alpha_{23}\sin(\frac{6\pi t}{12})+\beta_{23}\cos(\frac{6\pi t}{12})+\alpha_{24}\sin(\frac{8\pi t}{12})+\beta_{24}\cos(\frac{8\pi t}{12}) \quad (4)$$

最后，利用最小二乘法估计每个参数。虽然加进辅助回归元，但估计方法没有特殊变化，采用最小二乘法，化为不带差分项的简单形式的参数估计结果为：

$$Y_t=0.0099+0.9188Y_{t-1}-0.0553Y_{t-2}+0.354Y_{t-3}-0.2175Y_{t-4}+\varepsilon_t-0.8194\varepsilon_{t-1}+0.0211\sin(\frac{2\pi t}{2.6})-0.0072\cos(\frac{2\pi t}{2.6})-0.0347\sin(\frac{2\pi t}{12})-0.0875\cos(\frac{2\pi t}{12})+0.0948\sin(\frac{4\pi t}{12})-0.0502\cos(\frac{4\pi t}{12})+0.0036\sin(\frac{6\pi t}{12})+0.0244\cos(\frac{6\pi t}{12})+0.0221\sin(\frac{8\pi t}{12})-0.0506\cos(\frac{8\pi t}{12}) \quad (5)$$

利用估计的参数结果，将参数带到式(4)和式(5)，可以计算出周期为2.6个月和12个月的傅里叶级数序列，将序列的前24期画出折线图如图3所示：

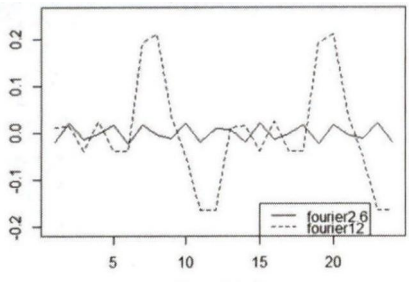

图3　傅立叶级数图

由图3可知，傅里叶级数的确较好地表示出了2.6个月和12个月的季节周期性，且周期为12个月的傅里叶级数也很好地体现了每年7-8月份客流量达到高峰，11-12月达到低谷的情况，因此将傅里叶级数与ARIMA模型部分结合，就可以很好地进行拟合和预测。

4　模型的检验与预测

首先，利用上文的估计结果对2004年1月至2015年8月的铁路客运量进行拟合，将拟合结果和原序列值绘制时间序列图，如图4所示：

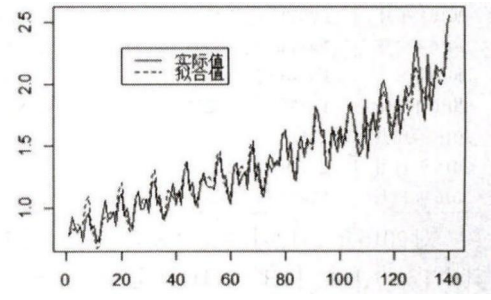

图4　铁路客运量拟合图

从图4可以看出拟合效果较好，趋势性和季节性都体现得比较明显，且平均绝对百分比误差(MAPE)为5.46%，说明效果很好。

再次，对模型的残差进行白噪声和正态性检验。首先画出残差序列图和正态QQ图，如图5和图6所示：

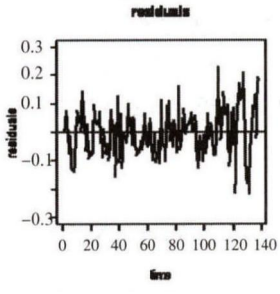

图5　残差序列图

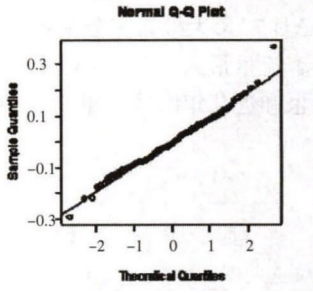

图6　残差QQ图

可见残差除了110,120,135期左右比较大外，基本没有规律性可言，对残差进行Ljung-Box检验结果显示LB统计量对应的P值为0.7965，大于0.05，说明残差确实是纯随机序列。

从QQ图可以看出，确实可以认为残差服从于正态分布，对残差进行Shapiro-Wilk正态性检验结果显示，W统计量对应的P值为0.1573>0.05，可以认为残差是服从于正态分布的。

从残差的白噪声和正态性检验可以看出模型是正确的。

最后，对2016年的客运量进行预测，预测首先需要构造傅里叶级数，从2015年9月到2016年12月共16个月，产生周期为2.6个月傅里叶级数的公式为：

$$y_t=-0.0211\sin(\frac{2\pi t}{2.6})+0.0072\cos(\frac{2\pi t}{2.6}),t=141,\cdots,156 \quad (6)$$

产生周期为12个月傅里叶级数的公式为：

$$y_t=-0.0347\sin(\frac{2\pi t}{12})-0.0875\cos(\frac{2\pi t}{12})+0.0948\sin(\frac{4\pi t}{12})-0.0502\cos(\frac{4\pi t}{12})+0.0036\sin(\frac{6\pi t}{12})+0.0244\cos(\frac{6\pi t}{12})+0.0221\sin(\frac{8\pi t}{12})-0.0506\cos(\frac{8\pi t}{12}),t=141,\cdots,156 \quad (7)$$

分别代入上述估计后的模型中，得到2016年12个月的预测值如表3所示：

表3　　2016年12个月的铁路客运量预测值　　（单位：亿人）

	预测值	Lo 80	Hi 80	Lo 95	Hi 95
2016年1月	2.2675	2.1364	2.3986	2.0670	2.4680
2016年2月	2.2560	2.1241	2.3879	2.0543	2.4578
2016年3月	2.2104	2.0755	2.3452	2.0042	2.4166
2016年4月	2.3051	2.1685	2.4417	2.0962	2.5140
2016年5月	2.2159	2.0779	2.3538	2.0049	2.4269
2016年6月	2.2632	2.1233	2.4031	2.0493	2.4772
2016年7月	2.4815	2.3399	2.6232	2.2649	2.6981
2016年8月	2.5046	2.3614	2.6477	2.2856	2.7235
2016年9月	2.3735	2.2286	2.5184	2.1519	2.5951
2016年10月	2.2532	2.1067	2.3997	2.0292	2.4773
2016年11月	2.1799	2.0318	2.3280	1.9534	2.4064
2016年12月	2.1874	2.0377	2.3371	1.9585	2.4163

对2016年1-12月的铁路客运量进行预测，结果显示，其中Lo 80、Lo95和Hi 80、Hi 95分别表示80%预测和95%预测区间的上下限，80%的置信区间比较窄，95%的置信区间有更大的可能性包含真值，铁路部门可以根据预测的置信区间的上限对运行的各方面进行规划，以保证铁路系统更顺利的运转，而不出现混乱。从表中可以看出1-3月中1月的客流量最大为2.268亿人，2月份也不低为2.256亿人，这个结果比较符合我国国情，因为2016年的除夕是2月7号，大批旅客会在1月末和2月初返乡；4-6月中4月客流量最大，其次是6月，但5月和6月客流量差不多，这也和近几年的情况相符，从2012年开始4到6月中5月的客流量最小，说明该模型能够较多的考虑时间点更近的信息；7-9月中8月的客流量最大为2.5046亿人，也为全年最高值，7月份在全年也较高客流量为2.482亿人，这个结果预测很符合我国8月份有大批学生返校和7-8月一定规模暑假游客增加的情况；10-12月中10月的客流量最大，十一国庆节放假七天，确实是外出高峰，说明预测结果较准确。预测中所有月份中11月和12月的客流量最小，这也是符合每年的情况的。

5　结论

运用傅立叶级数作为回归元的复杂季节ARIMA模型具有以下优点：①允许模型具有任何季节长度的周期。②允许模型包含多个季节周期。③从小变大的参数K可以描述从平缓到陡峭季节波动。④时间序列的短期波动可以通过ARIMA误差项来描述。

该方法在我国铁路客运量月度数据的应用中表现得较好，分别选择正余弦个数为1和4的2.6个月和12个月为周期的傅里叶级数作为辅助回归元拟合ARIMA(3,1,1)模型最优，其拟合的平均绝对百分比误差（MAPE）为5.46%。在此基础上对我国2016年各月份的客运量进行了科学预测，预测结果符合我国国情。

参考文献：

[1]Livera A M D, Hyndman R J, Snyder R D. Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing[J]. Journal of the American Statistical Association, 2011, 106(496).

[2]Gould P G, Koehler A B, Ord J K. Forecasting time Series With Multiple Seasonal Patterns[J].European Journal of Operational Research, 2008, 191(1).

（责任编辑/易永生）

Time Series Data With Complex Seasonal Periods Studies

Ma Jiayu,Han Zhaozhou

(College of Economics,Jinan University,Guangzhou 510632,China)

Abstract: Seasonal time series has more than one season cycle sometimes.For example website traffic data in hours, has a seasonal cycle with 24 hours a week and a seasonal cycle with a week. To solve the problem that traditional models can't handle complex seasons, this paper use Fourier series as the auxiliary regression of ARIMA model to fit China railway passenger traffic across 2004.1-2015.8 .The result shows that the best model is that use the Fourier series with the number of sin and cos are 1 and 3 and the cycle are 2.6 and 12 months to fit ARIMA (3,1,1) model, the MAPE is 5.46%. On the basis, we conduct a scientific prediction for the volume of railway passenger traffic of each months in 2016.

Key words:complex seasonality;fourier series;volume of railway passenger traffic