昆明医科大学学报 2012,(6):48~52
Journal of Kunming Medical University
CN 53-1049/R

ARIMA 乘积季节模型在食源性疾病月发病率预测中的应用

万 蓉，李娟娟，王晓雯
（云南省疾病预防控制中心，云南 昆明 650022）

［摘要］**目的** 探讨 ARIMA 乘积季节模型在食源性疾病发病率预测中的可行性，并预测食源性疾病的月发病率趋势. **方法** 对云南省 2004 年 1 月至 2010 年 12 月食源性疾病月发病率资料建立 ARIMA 乘积季节模型，利用 2011 年月发病率资料进行回代，预测 2012 年食源性疾病月发病率趋势. **结果** ARIMA(0，1，1)×(0，1，1)$_{12}$ 的 BIC 值最小，可以认为该模型的拟合优度相对最优；对该模型的残差进行白噪声检验，QLB（18）为 20.225（P=0.210），提示残差属于白噪声. **结论** ARIMA 乘积季节模型可以用于食源性疾病月发病率趋势的拟合和预测.

［关键词］食源性疾病；发病率； ARIMA 乘积季节模型；预测
［中图分类号］R155.3［文献标识码］A［文章编号］1003－4706（2012）06－0048－05

The Application of the SARIMA Model in Forcasting Month Incidence of Foodborne Diseases

WAN Rong，LI Juan－juan，WANG Xiao－wen
（Yunnan Center for Disease Control and Prevention，Kunming Yunnan 650022，China）

［**Abstract**］Objective To explore the accessibility of the SARMA model in forcasting the trend of the monthly incidence of foodborne disease. Methods We used the incidence dates from 2004 to 2011 to build up the SARIMA model and used the date in 2011 to confirm the model and forcast the monthly trend in 2012. Results ARIMA(0，1，1)×(0，1，1)$_{12}$ had the lest BIC value and had the best goodness of fit. By the white noise testing，QLB(18) = 20.225（P = 0.210），showing the residual error belonged to the white noisy. Conclusion SARIMA model can be used in the foodborne disease incidence fitting and forcasting.

［**Key words**］Foodborne disease；Incidence； SARIMA model；Forcasting

世界卫生组织在 1984 年将食源性疾病定义为通过摄食进入人体内的各种致病因子引起的、通常具有感染性质或中毒性质的一类疾病. 主要包括食物中毒、食源性肠道传染病、食源性寄生虫病、食源性化学物质污染食物引起的急、慢性中毒、食源性变态反应性疾病及食源性放射病[1].

对食源性疾病的监测和预测工作是疾病预防控制机构的一项重要职责，也是应急反应和处置的前提，其中，对食源性疾病发病率的预测能为采取主动预防措施提供科学的信息依据. 本文以云南省 2004 年 1 月至 2010 年 12 月食源性疾病月发病率为基础建立 ARIMA 乘积季节模型，利用 2011 年食源性疾病月发病率资料对模型进行回代，观察模型拟合效果，并预测 2012 年食源性疾病月发病率，从而为有效采取预防控制措施提供依据.

1 资料与方法

1.1 资料来源

资料来源于云南省各州/市、区/县疾病预防控制机构食源性疾病调查资料或统计分析报告，收集了 2004 年 1 月至 2011 年 12 月食源性疾病月发病人数，计算 2004 年 1 月至 2011 年 12 月食源性疾病月发病率（见表 1）.

［基金项目］科技部国家高技术研究发展计划（863 计划）资助项目（2010AA23004）
［作者简介］万蓉（1963～），女，重庆市人，学士，副主任医师，主要从事营养与食品卫生监测工作.
［通讯作者］王晓雯. E-mail:WXW_ph@163.com

表 1　2004 年至 2011 年食源性疾病月发病率（1/10 万）

Tab. 1 The month incidence of foodborn disease from 2004 to 2011（1/100 thousand）

年份	月份											
	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	0.00	0.32	0.26	0.05	0.73	0.46	0.24	0.61	0.25	0.22	0.28	0.00
2005	0.17	0.44	0.21	0.30	0.81	1.07	0.46	0.81	0.86	0.80	0.34	0.14
2006	0.31	0.16	0.46	0.16	0.02	0.23	0.66	0.64	0.50	0.32	0.16	0.30
2007	0.66	0.08	0.96	0.75	0.21	0.37	0.26	0.31	0.82	0.23	0.53	0.10
2008	0.03	0.16	0.01	0.75	0.79	0.33	1.01	0.02	0.69	0.20	0.42	0.17
2009	0.08	0.21	0.13	0.03	0.56	1.21	0.54	0.35	0.92	1.22	0.08	0.59
2010	0.12	0.02	0.303	0.74	0.12	0.30	0.06	0.87	0.12	0.48	0.23	0.20
2011	0.10	0.34	0.03	0.44	0.67	0.69	0.45	0.03	0.32	0.44	0.20	0.03

1.2　统计学处理

求和自回归滑动平均模型（ARIMA）[2-5]是 Box-Jenkins 方法中的重要时间序列分析预测模型. 该模型是用于描述非平稳资料的一种方法，主要包含 3 个过程：自回归、滑动平均和差分求和. 当时间序列资料含有季节性变动趋势时可以建立 ARIMA 乘积季节模型.

1.2.1　模型的识别　时间系列分析对时间序列数据有正态分布的假定，此外要求时间序列是平稳的. 如果一个时间序列的概率分布不随时间变化，为严格的平稳时间序列；如果时间序列的一、二阶矩存在，并在任意时刻均值是常数、协方差为时间间隔的函数，称为宽平稳时间序列，即通常所研究的时间序列. 非平稳时间序列资料，可以通过差分和数据变换的方式达到平稳化的目的. 通常采用图检验法、自（偏）相关函数检验法等进行简单判断. 根据自相关系数、偏自相关系数的结尾、拖尾性初步判断序列所适合的模型类型. 在确定模型后，采用自相关系数和偏自相关系数定阶法、最佳函数定阶法相结合的方法，识别模型阶数.

1.2.2　模型参数估计　常用的 ARIMA 模型参数估计方法包括矩估计法、极大似然估计法（MLE）、无条件最小二乘估计法（ULS）以及假定过去未观测的“三率”误差为 0 的条件最小二乘估计法（GLS）等.

1.2.3　模型的诊断　模型初步建立后，对模型进行诊断以选择最佳模型. ARIMA 模型要求残差为白噪声，即残差序列应服从正态分布，残差序列具有随机无趋势序列的自相关系数和偏自相关系数. 典型方法是通过拟合优度检验对观测值和模型拟合值的残差进行分析. 此外，BIC 准则考察模型对原始数据的拟合程度以选择最佳模型.

1.2.4　预报预测　应用所拟合的模型对 2011 年食源性疾病月发病率进行回代，观察模型拟合效果，并对未来 1 a 的月发病率进行预测，包括点估计值及其 95%的可信区间.

2　结果

2.1　模型的识别及初步拟合

绘制 2004 年 1 月至 2010 年 12 月食源性疾病月发病率直方图、序列图、自相关系数 ACF 图和偏自相关系数 PACF 图（见图 1），发现数据不呈正态分布，且存在明显的季节趋势，考虑先对数据进行自然对数变换，由于食源性疾病发病率数据中有 0 的存在，因此在进行对数变换前需要对数据进行调整，先把原序列中的每个值，加上一个很小的正值，该值为服从（0，0.1）均匀分布的随机数. 再将变换后的序列乘以一个大小在（0，1）之间的常数. 经过该变换后，序列的均值和标准差均不发生变化[5]. 随后进行一阶一般差分，再进行一阶季节差分分别消除趋势和季节的影响. 绘制平稳化后的 ACF 图和 PACF 图（见图 2），可见一阶一般差分和一阶季节差分后的序列平稳. 根据差分变换的阶数，可以确定模型形式为 SARIMA（p，1，q）×（P，1，Q）12，其中，p，q 和 P，Q 是待定的参数，分别表示连续性模型和季节模型中的自回归阶数和移动平均阶数. 将平稳化后的 ACF 图及 PACF 图与标准 ARIMA 图比较[2]，初步判断连续模型为 ARIMA（0，1，1）. 季节模型的参数 P、Q 判断较难，但根据文献，参数超过 2 的情况很少见[6]，可以分别取 0、1、2 由低阶到高阶逐个试验，根据模型的拟合优度、残差情况以及系数间的相关性进行综合判断.

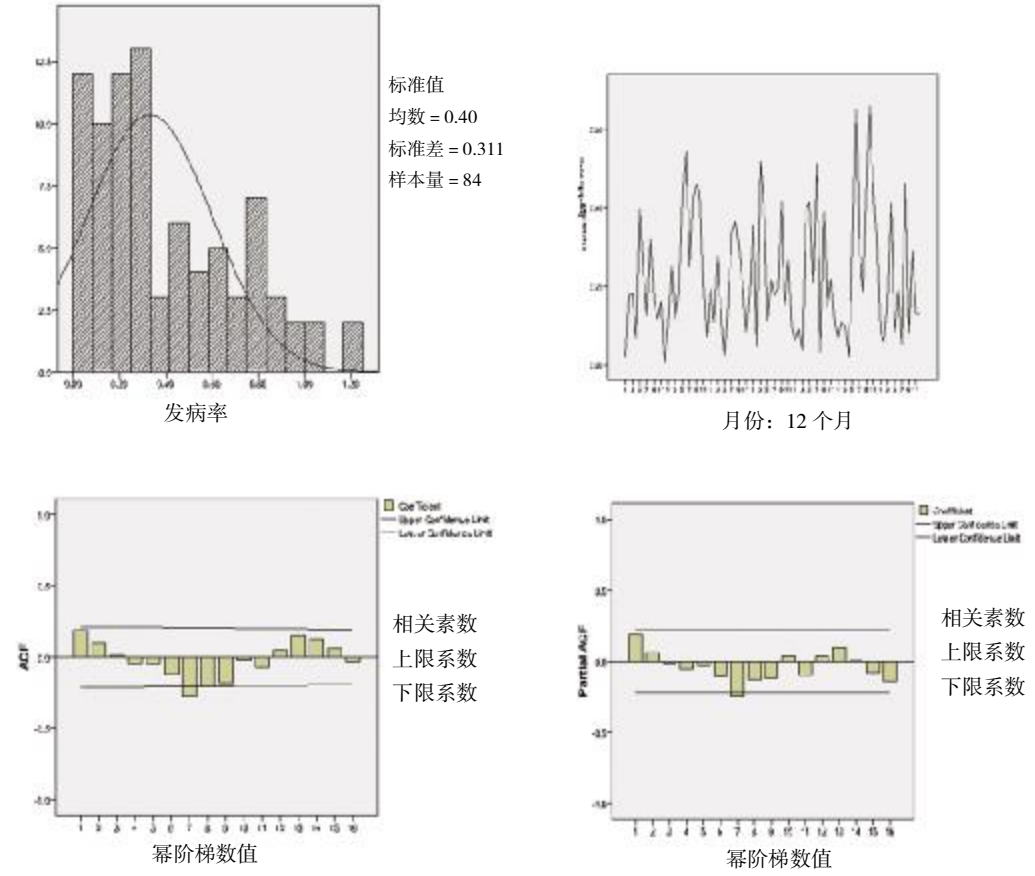

图 1　2004 年至 2010 年食源性疾病月发病率直方图、序列图、ACF 图和 PACF 图

Fig. 1　The histogram, sequence map, ACF and PACF map of the month incidence of foodborn disease from 2004 to 2011

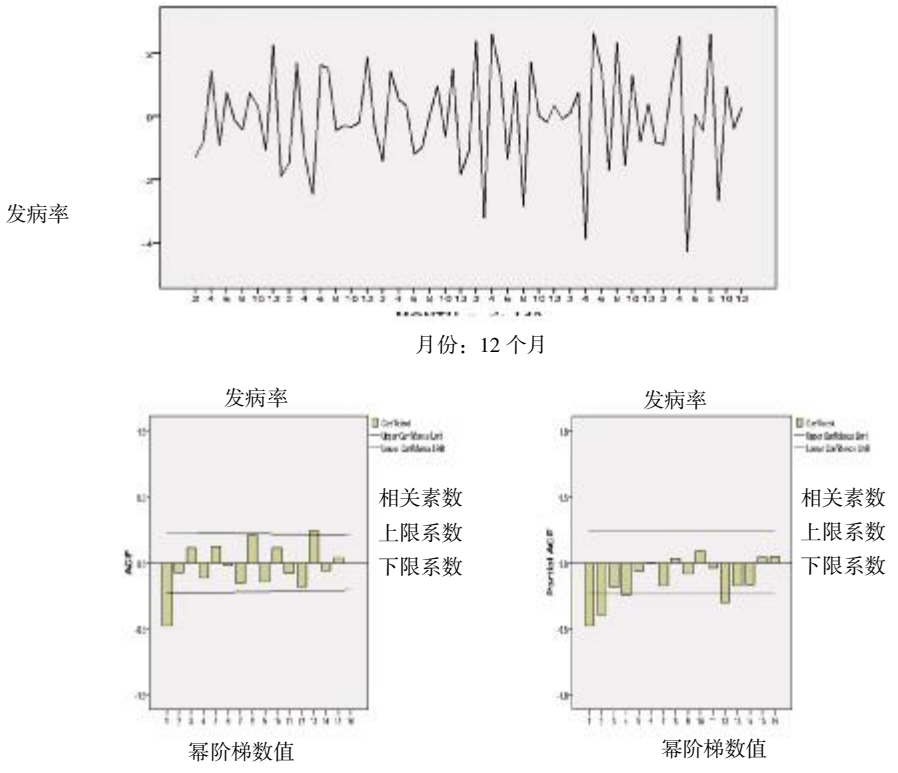

图 2　原系列经平稳化后的序列图、ACF 图和 PACF 图

Fig. 2　The sequence map, ACF and PACF map of the original data after stabilizing

2.2　模型的参数估计与模型诊断

备选模型的参数估计见表 2，模型的诊断从以下几个方面进行：（1）备选模型的 R2 和 Normalized BIC 值进行比较发现，表 2 显示 ARIMA(0,1,1)×(0，1，2)$_{12}$ 的 BIC 值最小，因此可以认为该模型的拟合优度相对最优；（2）对该模型的残差进行白噪声检验，QLB（18）12.198（P=0.664），提示残差属于白噪声，表明该模型对食源性疾病月发病率序列所包含的信息提取完全[3]，说明所选模型是恰当的.

用本模型预测 2011 年食源性疾病月发病率结果见表 3 所示. 可以看出模型预测值的动态趋势与实际情况基本一致. 2011 年各月的月发病率虽然与预测值不完全一样，但各月预测值的 95%的可信区间范围内均包括了实测值，因此，可以认为模型对未来情况可以进行很好的追踪和预测.

表 2　备选模型参数估计

Tab. 2　The parameter estimation of substituted models

参　数	ARIMA(0,1,1)×(0,1,1)				ARIMA(0,1,1)×(0,1,2)				ARIMA(0,1,2)×(0,1,1)			
	B	SEB	t	P	B	SEB	t	P	B	SEB	t	P
常数	−0.004	0.005	−0.845	0.401	−0.008	0.006	−1.317	0.192	−0.004	0.005	−0.848	0.400
MA1	0.990	0.470	2.106	0.039	0.999	5.775	0.173	0.863	1.008	0.751	1.342	0.184
MA2	–	–	–	–	–	–	–	–	−0.014	0.131	−0.107	0.915
SMA1	0.881	0.374	2.353	0.022	0.636	636.0	0.001	0.999	0.882	0.379	2.329	0.023
SMA2	–	–	–	–	0.363	363.0	0.001	0.999	–	–	–	–

表 3　备选模型拟合优度统计量

Tab. 3　The statistics of goodness of fit of substituted models

参　数	ARIMA(0,1,1)×(0,1,1)	ARIMA(0,1,1)×(0,1,2)	ARIMA(0,1,2)×(0,1,1)
R2	0.550	0.556	0.550
Normalized BIC	−0.680	−0.953	−0.543

表 4　食源性疾病 2011 年月发病率实际值与预测值比较

Tab. 4　Comparison between the actual value and prognostic value the month incidence of foodborn disease in 2011

发病率	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
实际值	0.10	0.34	0.03	0.44	0.67	0.69	0.45	0.03	0.32	0.44	0.20	0.03
预测值	0.11	0.10	0.18	0.35	0.14	0.22	0.13	0.29	0.16	0.21	0.14	0.07
CL 上限	0.46	0.45	0.77	1.55	0.76	0.98	0.55	1.29	0.70	0.93	0.62	0.32
CL 下限	0.01	0.01	0.02	0.03	0.01	0.02	0.01	0.02	0.01	0.02	0.01	0.01

注：CL:95%可信区间.

2.3　模型的预测

用所建立的模型外推 2012 年食源性疾病的月发病率，见图 3. 提示 2012 年云南省食源性疾病的月发病率水平基本稳定，并呈现整体下降的趋势.

3　讨论

云南省食源性疾病的发病呈现出明显的季节高峰，通常在第 2、3 季度到达高峰[7]，此次收集的食源性疾病月发病率资料显示了这种季节趋势

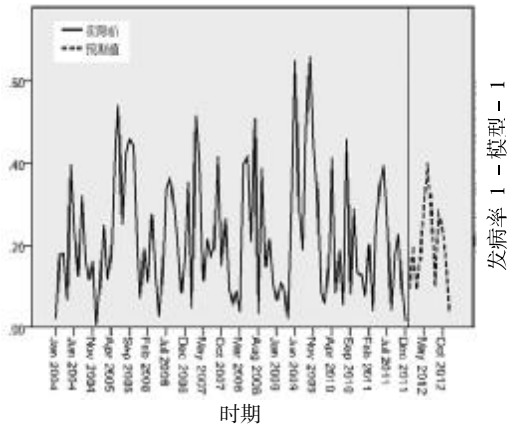

图 3　ARIMA(0，1，1)×(0，1，1)$_{12}$ 模型 2012 年预测值

Fig. 3　The prognostic value of ARIMA(0,1,1)×(0,1,1)$_{12}$ model in 2012

的存在. 吴家兵等[8]探讨了ARIMA季节乘积模型在传染病发病率预测中的应用，并指出了ARIMA模型在传染病预测中简便、适用和短期预测精度较高等优势. 本文将ARIMA模型运用于同样具有趋势性、季节性和周期性特点的食源性疾病发病率的预测. 考虑到食源性疾病发病率明显的季节性，建立了ARIMA季节乘积模型，经过模型诊断发现所选定的模型对食源性疾病发病率序列所包含的信息提取完全，可以用于食源性疾病月发病率的预测. 通过模型回代，发现2011年的发病率预测值与实际值基本吻合. 因此，可以认为ARIMA季节乘积模型适用于食源性疾病月发病率的预测. 在今后的研究中，应当首先确保所收集的数据的准确性，同时，应当收集相关新观察数据并对数据重新进行模型拟合，保证和提高预测效果.

在ARIMA模型的实际应用中，应注意ARIMA模型的应用前提是事件序列的平稳性，实际工作中数据往往是非平稳序列，需要对序列进行预处理，使之达到平稳化的要求；同时，序列不能太短，会影响到预测的可靠性. 此外，如果研究对象的趋势发生了较大的改变，则需要积累新的数据实时对型进行修正和重新拟合.

[参考文献]

[1] 黄承钰. 医学营养学[M]. 北京：人民卫生出版社，2003：193.

[2] 孙振球. 医学统计学[M]. 北京：人民卫生出版社，2002：461－477.

[3] 朋文佳，朱玉，何倩，等. ARIMA乘积季节模型在细菌性痢疾月发病率预测中的应用[J]. 中国卫生统计，2011，28(6)：645－647.

[4] 彭志行，陶红，贾成梅，等. 时间序列分析在麻疹疫情预测预警中的应用研究[J]. 中国卫生统计，2010，27(5)：459－462.

[5] AYAKO SUMI，KEN-ICHI KAMO，NORIO OHTOMO，et al. Time series analysis of incidence data of influenza in japan[J]. Epidemiology，2012，21(1)：21-29.

[6] 温亮，徐德忠，林明和，等. 应用实践序列模型预测疟区疟疾发病率[J]. 第四军医大学学报，2004，25(6)：507-510.

[7] 赵江，万蓉. 2008年至2009年云南省食物中毒流行特征分析[J]. 中国公共卫生管理，2011，27(1)：98－99.

[8] 吴家兵，叶临湘，尤尔科. ARIMA模型在传染病发病率预测中的应用[J]. 数理医学杂志，2007，20(1)：90－92.

（2012－03－10收稿）