第二军医大学学报 2013年9月第34卷第9期 http://www.ajsmmu.cn
Academic Journal of Second Military Medical University, Sep. 2013, Vol. 34, No. 9

DOI:10.3724/SP.J.1008.2013.00980

·论 著·

ARIMA模型在涂阳肺结核月发病预测中的应用

杨海琴[1],胡代玉[2],刘 瑛[2],王润华[1], 易 静[1*]
1.重庆医科大学公共卫生与管理学院卫生统计与信息管理学教研室,重庆 400016
2.重庆市结核病防治所,重庆 400020

[摘要] **目的** 探讨重庆市涂阳肺结核月发病数随时间的变化规律,为控制和预防肺结核提供科学依据。**方法** 采用SPSS13.0软件对2005～2009年重庆市涂阳肺结核月发病数资料建立ARIMA模型,利用该模型预测2010年1月～12月的涂阳肺结核月发病数,对模型的短期预测及其效果进行初步评价。**结果** 建立的ARIMA$(1,1,0)\times(0,1,1)_{12}$模型是拟合重庆市涂阳肺结核月发病数的合适模型,2005～2009年观测值落在拟合值95%的可信区间内,2010年预测值的平均相对误差为6.31%。**结论** ARIMA$(1,1,0)\times(0,1,1)_{12}$模型能很好地预测重庆市涂阳肺结核月发病情况,为控制和预防肺结核提供了可靠依据。

[关键词] ARIMA模型; 肺结核; 预测; 发病率
[中图分类号] R 521 [文献标志码] A [文章编号] 0258-879X(2013)09-0980-05

Application of ARIMA model in forecasting monthly incidence of smear-positive tuberculosis

YANG Hai-qin[1], HU Dai-yu[2], LIU Ying[2], WANG Run-hua[1], YI Jing[1*]
1. Department of Health Statistics and Information Management, College of Public Health and Management, Chongqing Medical University, Chongqing 400016, China
2. Department of Chongqing Tuberculosis Prevention, Chongqing 400020, China

[**Abstract**] **Objective** To investigate the variation of the monthly incidence of smear-positive tuberculosis with time in Chongqing, so to provide a scientific evidence for the control and prevention of tuberculosis. **Methods** Using the SPSS 13.0 software, we established an ARIMA model with the monthly incidence data of smear-positive tuberculosis (2005-2009), and the model was used to forecast the monthly incidence of Jan. 2010 to Dec. 2010. The short-term forecasting efficacy was evaluated. **Results** The established ARIMA $(1,1,0)\times(0,1,1)_{12}$ model was suitable for forecasting the monthly incidence of smear-positive tuberculosis in Chongqing. The observed values of 2005-2009 were in the 95% confidence interval of the fitted values, and the average relative error of the predictive value was 6.31% for 2010. **Conclusion** ARIMA $(1,1,0)\times(0,1,1)_{12}$ model can satisfactorily forecast the monthly incidence of smear-positive tuberculosis in Chongqing, which provides a reliable evidence for control and prevention of tuberculosis.

[**Key words**] ARIMA model; pulmonary tuberculosis; forecasting; incidence
[Acad J Sec Mil Med Univ, 2013, 34(9): 980-984]

结核病是由结核分枝杆菌引起的严重危害人类健康的慢性传染病,是世界卫生组织和我国重点控制的传染病之一。据世界卫生组织统计,我国是全球22个肺结核病流行严重的国家之一,同时也是全球27个耐多药肺结核病流行严重的国家之一[1]。国家卫生部于2010年举行的第5次全国肺结核流行病学抽样调查结果显示,全国肺结核患病率继续呈现下降趋势,防治工作取得显著效果。但是,当前我国肺结核病疫情形势依然严峻,防治工作仍面临诸多挑战。预测涂阳肺结核的发病情况是防治涂阳肺结核工作中一个非常重要的环节,根据涂阳肺结核发病的变化规律建立涂阳肺结核病疫情的预测模型,有效地预测涂

[收稿日期] 2013-03-01 [接受日期] 2013-08-02
[基金项目] 国家自然科学基金(30872160),重庆市科委自然科学基金(CSTC,2009BB5415). Supported by National Natural Science Foundation of China (30872160) and Natural Science Foundation of Chongqing Science and Technology Committee (CSTC,2009BB5415).
[作者简介] 杨海琴,硕士生. E-mail: 18716693627@163.com
* 通信作者(Corresponding author). Tel: 023-88369772, E-mail: yijinga@sina.com

阳肺结核的发病人数,对预防和治疗肺结核病有重大意义。本研究采用时间序列分析方法中的乘积季节 ARIMA 模型对重庆市 2005 年 1 月～2010 年 12 月的涂阳肺结核发病例数进行拟合和预测,探讨模型的可行性,为检测、预防和治疗肺结核提供科学依据。

1 材料和方法

1.1 *数据来源* 资料来源于重庆市结核病防治所提供的 2005～2010 年重庆市涂阳肺结核月发病人数,包含了重庆市所有区、县的涂阳肺结核人数,全面掌握了整个地区涂阳肺结核的发病人数,具有良好的代表性。

1.2 *ARIMA 模型的建立*

1.2.1 *ARIMA 模型的理论基础* 20 世纪 60 年代,美国学者 Box 和英国统计学者 Jenkins 提出了一整套关于时间序列分析、预测和控制的方法,被称为 Box-Jenkins 建模方法[2]。ARIMA(自回归综合移动平均)模型是时间序列分析中最常用的模型,也被称之为 Box-Jenkins 模型,或带差分的自回归移动平均模型。按照模型包含季节性成分与否,模型可分为季节性 ARIMA 模型 $(P,D,Q)_s$、非季节性 ARIMA 模型(p,d,q)和 ARIMA 混合模型 $(p,d,q)\times(P,D,Q)_s$,其中 p、d、q 和 P、D、Q 分别为非季节性和季节性自回归(AR)、差分(I)、移动平均(MA)的阶数,s 为季节周期因子。

1.2.2 *ARIMA 模型的建立步骤* 建立 ARIMA 模型通过以下 4 个步骤:序列的平稳化、模型定阶、估计参数和诊断模型。

(1)替换缺失值:时间序列模型一般都要求序列数据完整无缺,但实际上数据常是不完整的。当序列中存在缺失数据时,如果采用直接剔除数据的方法,容易使剔除缺失值之后的数据周期发生错位,而无法得到正确的分析结果。在这种情况下,应当使用替换缺失值过程,采用适当的方法对缺失值进行替换。在时间序列分析中,常用的缺失值替换方法有序列均值法、临近点均值法、临近点中位数法、线性插值法、线性回归法等。根据研究目的和时间序列的具体情况,选择合适的替换缺失值的方法。

(2)序列平稳化:建立 ARIMA 模型的前提是时间序列是平稳的,就是要求时间序列满足以下 3 个条件:序列的均值不随时间变化;序列的方差不随时间变化;序列的自相关系数只与时间间隔有关,而与所处的具体时刻无关。实际上,用初始数据建立的时间序列大多是不平稳的。当序列不平稳时,要通过平稳化过程将其转为平稳的序列。平稳化分为方差平稳化和趋势平稳化,前者通过平方根或自然对数转换实现,后者通过非季节性差分或季节性差分实现。

(3)模型的定阶:作时间序列的自相关图和偏自相关图,观察其自相关系数(ACF)和偏自相关系数(PACF),初步确定模型的阶数,即 p、d、q、P、D、Q、s 的值。

(4)估计参数和诊断模型:通过非线性最小二乘回归法或最大似然法估计模型的系数,并检验其显著性。参数估计后,检验模型的残差序列是否为白噪声,用以判断所建立模型的适合性。一个适合的模型的残差序列应是白噪声过程,其 ACF 和 PACF 不应与 0 有显著性差异; Box-Ljung Q 统计量应无显著性[3]。

1.3 *统计学处理* 采用 SPSS 13.0 软件对数据进行处理和分析,以 2005 年 1 月～2009 年 12 月的涂阳肺结核月发病数建立 ARIMA 模型,利用 2010 年 1 月～12 月的涂阳肺结核月发病数检验模型的预测效果。

2 结 果

2.1 *缺失值的替换* 本研究的数据中,2008 年 10 月～12 月 3 个月的涂阳肺结核病例数是缺失的。采用缺失值替换常用的方法是序列均值法对序列中的缺失值进行填补,2008 年 10 月～12 月的替换值分别为 1 168、1 168、1 168。

2.2 *序列的平稳化* 绘制重庆市 2005 年 1 月至 2009 年 12 月涂阳肺结核月发病例数的时序图(图 1)。时序图显示该序列的方差波动比较大,发病高峰和低峰之间的间距比较大,揭示了原始序列不是平稳的。另外,从时序图中可以看出,该序列的周期性比较明显,其中以 3 月份报告发病例数最多,12 月至次年 1 月份发病例数比较低。

为了消除序列的方差波动比较大,用常用的平方根或自然对数转换进行试验,发现原始序列经过自然对数转换后方差的波动明显减小。由于该序列具有明显的周期性,通过一次季节性差分消除其周期性,绘制其 ACF 图和 PACF 图(图 2)。结果显示,自相关系数和偏自相关系数都在滞后 12 期后不明显,说明存在非季节性成分。对经过上述转换后的序列再进行 1 次差分转换,并绘制其 ACF 图和 PACF 图(图 3),结果显示经过转换后的序列较平稳。

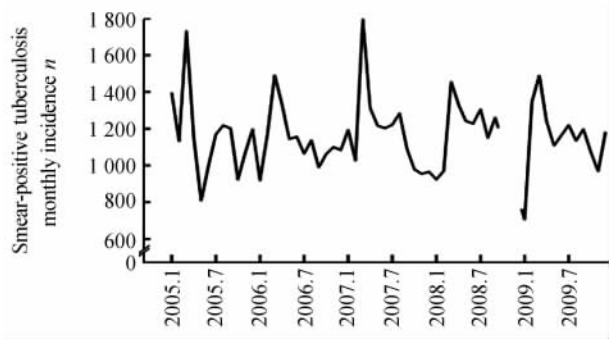

图 1　重庆市 2005～2009 年涂阳肺结核月发病数时序图

Fig 1　Time sequence of monthly incidence of smear-positive tuberculosis in Chongqing during 2005-2009

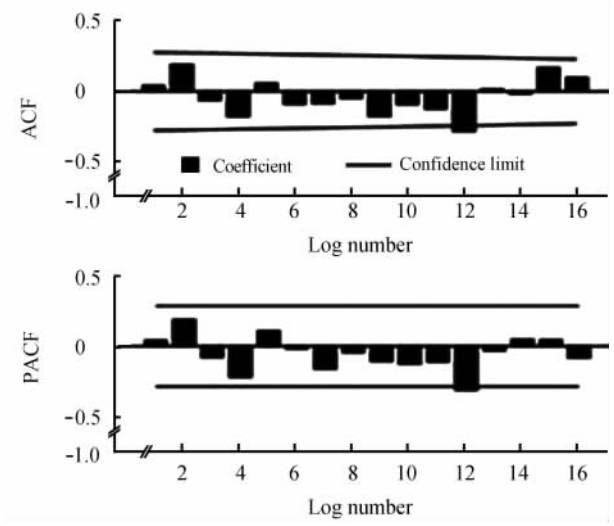

图 2　原序列经过自然对数转换和一次季节差分转换后的 ACF 和 PACF 图

Fig 2　ACF and PACF plots of original sequence after the conversion of natural logarithm and once the season difference

2.3　*模型的定阶*　模型的定阶就是确定 ARIMA 模型中各个参数的值。原始序列经过自然对数转换、一次非季节差分和一次季节差分转换之后的序列是平稳的，所以 s＝1，d＝D＝1。图 3 显示，自相关系数和偏相关系数都在时滞为 1 处显著不为 0，在其他时滞处与 0 的差异无统计学意义，可以初步确定 P＝q＝1。季节模型的参数 P、Q 判断较困难，但根据文献，参数超过 2 的情况很少，可以从低阶到高阶反复试验，根据模型的拟合优度、残差情况及系数间的相关性进行综合判断[4]。

2.4　*估计参数*　通过上面的步骤确定备选模型是 ARIMA(1,1,0)×(0,1,1)$_{12}$，ARIMA(1,1,1)×(1,1,0)$_{12}$和 ARIMA(1,1,1)×(0,1,1)$_{12}$。各备选模型的参数估计和拟合优度统计量分别见表 1、表 2。

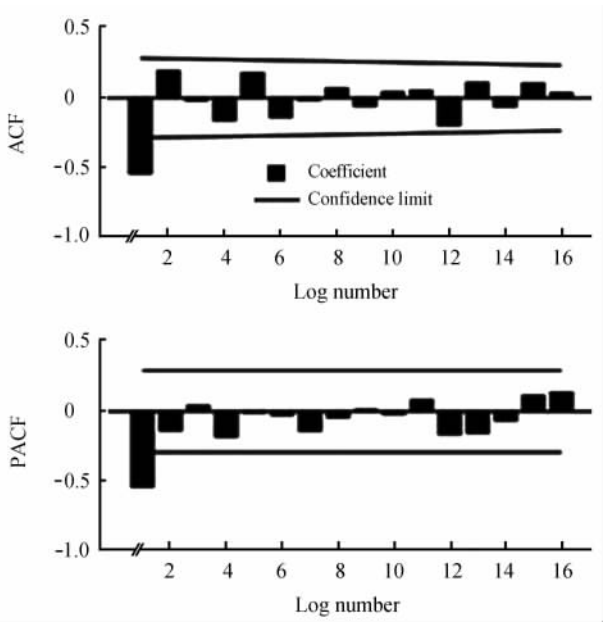

图 3　原序列经过自然对数转换、一次季节差分和一次非季节差分转换后的 ACF 和 PACF 图

Fig 3　ACF and PACF plots of original sequence after the conversion of natural logarithm, a season difference and a non-season difference

判断模型的优劣常用的方法有 BIC 准则和 AIC 准则，AIC 综合考虑似然函数和参数个数，而 BIC 比 AIC 多考虑了残差个数的影响，是 1 个通用的标准[5]。从表 3 可以看出，3 个备选模型的 AIC 值和 BIC 值都差不多，ARIMA(1,1,1)×(1,1,0)$_{12}$和 ARIMA(1,1,1)×(0,1,1)$_{12}$模型的 AIC 值和 BIC 值相对较小。但是从表 2 可以看出，只有模型 ARIMA(1,1,0)×(0,1,1)$_{12}$的各个参数估计的 $P<0.05$，参数具有统计学意义，且各参数之间相关性($r=0.022$)较小。因此选用 ARIMA(1,1,0)×(0,1,1)$_{12}$模型进行预测。

2.5　*模型的诊断*　模型诊断的另外一个重要手段就是对模型拟合后残差的独立性进行检验，即白噪声检验。图 4 是 ARIMA(1,1,0)×(0,1,1)$_{12}$模型残差的独立性检验结果，残差的 Box-Ljung 统计结果显示参数无统计学意义(表 3)，表明残差序列是白噪声。因此，可认为 ARIMA(1,1,0)×(0,1,1)$_{12}$模型是比较理想和简约的模型。

2.6　*模型的预测及评价*　根据所建的模型将 2005 年 1 月至 2009 年 12 月的涂阳肺结核发病数进行组内回代预测，并对 2010 年 1 月～12 月发病数进行组外预测(图 5)。结果表明：涂阳肺结核月发病数的组内回代预测数据与实际数据基本吻合，且均落入 95%

可信区间内。时间序列分析的主要目的是预测未来值并评估其变化趋势。本研究对 2010 年 1 月～12 月涂阳肺结核发病数进行预测，结果显示观测值与预测值之间平均相对误差为 6.31%(表 4)。

表 1 各备选 ARIMA 模型的参数估计值

Tab 1 Parameter estimates of all alternative ARIMA models

Parameter	$ARIMA(1,1,0)\times(0,1,1)_{12}$			$ARIMA(1,1,1)\times(1,1,0)_{12}$			$ARIMA(1,1,1)\times(0,1,1)_{12}$		
	Estimated value	t value	P value	Estimated value	t value	P value	Estimated value	t value	P value
AR(1)	−0.58	−4.72	0.00	−0.32	−1.46	0.15	−0.20	−0.90	0.37
MA(1)	—	—	—	0.46	2.33	0.02	0.55	2.93	0.00
SAR(1)	—	—	—	−0.48	−2.97	0.00	—	—	—
SMA(1)	0.51	2.14	0.04	—	—	—	0.48	2.14	0.04
Constant	0.00	0.44	0.66	0.00	0.31	0.76	0.00	0.37	0.71

表 2 各备选 ARIMA 模型的拟合优度统计量

Tab 2 Goodness-of-fit statistics of all alternative ARIMA models

Statistics	Model 1	Model 2	Model 3
Std	0.16	0.16	0.16
Log-likelihood values	18.59	19.51	19.54
AIC	−31.19	−31.03	−31.07
BIC	−25.64	23.63	−23.67

Model 1: $ARIMA\ (1,1,0)\times(0,1,1)_{12}$; Model 2: $ARIMA\ (1,1,1)\times(1,1,0)_{12}$; Model 3: $ARIMA\ (1,1,1)\times(0,1,1)_{12}$. Std: Model Std. error; AIC: Akaike's information criterion; BIC: Schwarz's Bayesian criterion

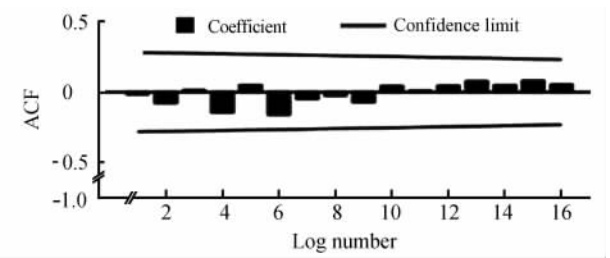

图 4 $ARIMA(1,1,0)\times(0,1,1)_{12}$模型残差序列的 ACF 图

Fig 4 ACF plots of residual series of $ARIMA(1,1,0)\times(0,1,1)_{12}$ model

表 3 $ARIMA(1,1,0)\times(0,1,1)_{12}$模型自相关系数及检验

Tab 3 Autocorrelation coefficient and examination of $ARIMA\ (1,1,0)\times(0,1,1)_{12}$ model

Lag	Autocorrelation coefficient	Std	Box-Ljung statistics	
			Value	P value
1	−0.02	0.14	0.01	0.91
2	−0.08	0.14	0.32	0.85
3	0.02	0.14	0.33	0.95
4	−0.14	0.14	1.46	0.83
5	0.05	0.14	1.60	0.90
6	−0.16	0.13	3.05	0.80
7	−0.05	0.13	3.18	0.87
8	−0.02	0.13	3.22	0.92
9	−0.07	0.13	3.53	0.94
10	0.04	0.13	3.65	0.96
11	0.01	0.12	3.66	0.98
12	0.04	0.12	3.79	0.98
13	0.08	0.12	4.20	0.99
14	0.05	0.12	4.37	0.99
15	0.08	0.12	4.83	0.99
16	0.05	0.12	5.04	1.00

3 讨 论

时间序列分析方法是依据历史数据随时间发生变化的规律，建立时序模型，以达到预测未来的目的。ARIMA 模型是时间序列分析中最常用的方法之一，该方法不需要对时间序列的发展模式作先验假设，可通过反复识别、估计、诊断，获得合适的结果，此过程借助于计算机操作，是一种简单方便、实用性强、精确度较高的短期预测方法[6]。近年来，该方法已经广泛应用于医学领域各个方面，特别是传染病的发病或死亡的预测预报工作[6-9]。

近年来，ARIMA 模型已经广泛应用于肺结核发病率的预测与检测[10]。本研究利用重庆市 2005 年 1 月至 2009 年 12 月涂阳肺结核月发病数资料，建立了 $ARIMA(1,1,0)\times(0,1,1)_{12}$预测模型，预测结果显示，涂阳肺结核月发病数实际值都落入预测值 95%可信区间内，预测值随时间变化的规律与实际情况基本吻合，平均相对误差小于 10%，表明利用 $ARIMA(1,1,0)\times(0,1,1)_{12}$模型预测重庆市涂阳肺结核发病趋势的可行性。另一方面，也显示了预测的应用价值和实用性。$ARIMA(1,1,0)\times(0,1,1)_{12}$模型可以预测重庆市涂阳肺结核的发病数，从而预测涂阳肺结核的发病趋势，优化肺结核的预防和早期预警系统。依据早期预警系统，可以采取社区干预和个人防护等措施预防肺结核。

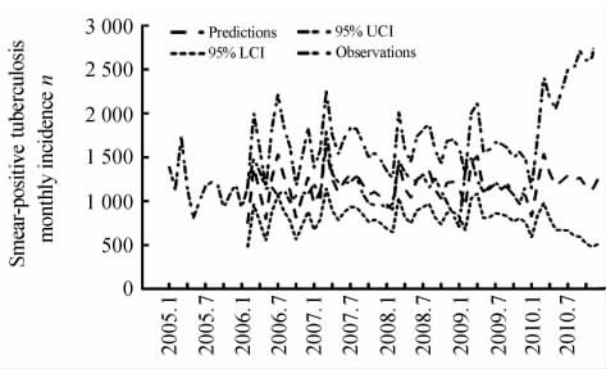

图 5　重庆市 2005 年 1 月～2010 年 12 月涂阳肺结核月发病数序列预测图

Fig 5　Sequence prediction plots of smear-positive tuberculosis monthly incidence from Jan. 2005 to Dec. 2010 in Chongqing

表 4　2010 年重庆市涂阳肺结核发病数预测及评价结果

Tab 4　Incidence prediction and evaluation of Chongqing smear-positive tuberculosis in 2010

Month	Incidence *n*		Error	
	Observations	Predictions	Absolute error	Relative error(%)
1	903	822.75	−80.25	−8.89
2	1 159	1 212.95	53.95	4.66
3	1 560	1 539.18	−20.82	−1.33
4	1 224	1 319.19	95.19	7.78
5	1 189	1 176.69	−12.31	1.04
6	1 123	1 232.93	109.93	9.79
7	1 185	1 292.91	107.91	9.11
8	1 146	1 237.45	91.45	7.98
9	1 159	1 269.34	110.34	9.52
10	1 064	1 163.71	99.71	9.37
11	1 095	1 124.38	29.38	2.68
12	1 213	1 256.26	43.26	3.57
Mean	—	—	71.21	6.31
Total	14 020	14 647.75	627.75	4.48

ARIMA 模型已经被证实了能广泛应用于传染病的发病预测[11]。本研究证明了 ARIMA 模型能较好地预测涂阳肺结核的发病趋势。由于不同地区、不同时间，即使是同一种传染病，其发生发展的规律也不尽相同，构建的模型也不一定相同。对于单次分析所建立的模型，不能作为长期预测的工具，该模型只能用于短期预测。要随着时间的推移，不断地注入新的数据，重新建立模型，才能及时有效地预测传染病的发病趋势。使用 ARIMA 模型进行预测，如果对研究对象采取了预防接种或加强环境治理等干预措施后，此时应该结合实际情况，全面考虑慎重使用预测结果，并且需要引入新的数据重新拟合模型，方可达到可靠预测。本研究只关注了涂阳肺结核病例数，没有考虑肺结核发生的因素，如天气、生物、社会和经济因素等，研究结果难免存有偏倚，仍有待进一步研究校正。

4　利益冲突

所有作者声明本文不涉及任何利益冲突。

[参 考 文 献]

[1] 中华人民共和国卫生部. 卫生部介绍全国肺结核疫情现状[EB/OL]. 2011. *http://www.moh.gov.cn./publicfiles/business/htmlfiles/mohjbyfkzj/s3590/201103/51027.htm*

[2] Box G E, Jenkins G M. 时间序列分析预测与控制[M]. 北京：中国统计出版社，1997：500.

[3] 张彦琦，唐贵立，王文昌，易　东. ARIMA 模型及其在肺结核预测中的应用[J]. 现代预防医学，2008，35：1608-1610.

[4] 万　蓉，李娟娟，王晓雯. ARIMA 乘积季节模型在食源性疾病月发病率预测中的应用[J]. 昆明医科大学学报，2012，6：48-52.

[5] 朱继明，汤林华，周永森，黄　芳. 非稳定性疟区用时间序列模型预测疟疾发病率的可行性研究[J]. 中国寄生虫病学与寄生虫杂志，2007，25：232-235.

[6] 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]. 数理医药学杂志，2007，20：90-92.

[7] Earnest A, Tan S B, Wilder-Smith A, Machin D. Comparing statistical models to predict dengue fever notifications[J]. Comput Math Methods Med, 2012, 2012: 758674.

[8] Hanf M, Adenis A, Nacher M, Carme B. The role of El Niño Southern Oscillation (ENSO) on variations of monthly *Plasmodium falciparum* malaria cases at the Cayenne General Hospital, 1996-2009, French Guiana [J]. Malar J, 2011, 10: 100.

[9] Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White N J, Kaewkungwal J. Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan[J]. Malar J, 2010, 9: 251.

[10] Yi J, Du C T, Wang R H, Liu L. Applications of multiple seasonal autoregressive integrated moving average (ARIMA) model on predictive incidence of tuberculosis[J]. Zhonghua Yu Fang Yi Xue Za Zhi, 2007, 41: 118-121.

[11] Lin H, Yang L, Liu Q, Wang T, Hossain S R, Ho S C, et al. Time series analysis of Japanese encephalitis and weather in Linyi City, China[J]. Int J Public Health, 2012, 57: 289-296.

[本文编辑]　贾泽军