

应用 ARIMA 模型对麻疹发病的预测及分析

梁静, 王培生, 李新风, 芮宝玲, 陈保林

乌鲁木齐市疾病预防控制中心, 新疆 乌鲁木齐 830026

摘要: **目的** 应用 ARIMA 模型对麻疹发病数预测并探讨其可行性, 为防控麻疹疫情提供依据。 **方法** 采用 SPSS17.0 对乌鲁木齐市 2009–2015 年麻疹月发病数的资料建立 ARIMA 模型, 并预测 2016 年麻疹月发病数。 **结果** 建立模型 $ARIMA(0,0,1)(1,1,0)_{12}$ 是合适的, 且模型检验自相关系数在 ± 0.5 之间, 预测值与实际值有较高的吻合度。 **结论** ARIMA 模型能很好的拟合乌鲁木齐市麻疹发病数趋势, 预测效果可靠。

关键词: 麻疹; 时间序列; ARIMA 模型; 预测

中图分类号: R511.1 **文献标识码:** B **文章编号:** 1006–3110(2016)08–1003–04 **DOI:** 10.3969/j.issn.1006–3110.2016.08.035

Application of ARIMA model to forecast and analysis of measles incidence

LIANG Jing, WANG Pei-sheng, LI Xin-feng, RUI Bao-ling, CHEN Bao-lin

Urumchi Municipal Center for Disease Control and Prevention, Urumchi, Xinjiang 830026, China

Abstract: **Objective** To forecast the measles incidence by ARIMA model, to verify its practicability, and to provide evidence for prevention and control of measles epidemic. **Methods** SPSS17.0 software was used to establish ARIMA model about the monthly incidence of measles in Urumchi during 2009–2015, and then the monthly incidence of measles in 2016 was forecasted.

Results The established $ARIMA(0,0,1)(1,1,0)_{12}$ model was satisfactory and practicable, and the correlation coefficient of model test was within ± 0.5 . The predicted values agreed well with the actual values. **Conclusions** It is practical to apply the approach of ARIMA model to predict the incidence of measles in Urumchi.

Key word: Measles; Time-series; ARIMA mode; Forecast

在应用麻疹减毒活疫苗(measles attenuated live vaccine, MV)前, 中国每隔一年周期性出现一次麻疹流行高峰, 随着 MV 常规免疫的实施, 全国报告麻疹发病水平持续下降^[1], 2012 年全国麻疹发病率降到 0.46/10 万。乌鲁木齐市 2012 年麻疹发病率为 5.48/10 万, 观察近十年数据发现乌鲁木齐市麻疹发病率均高于全国和新疆的发病率, 这与 2013 年国务院《卫生事业发展“十二五”规划》提出要实现消除麻疹的目标^[2]相距甚远。为分析乌鲁木齐市麻疹的流行趋势, 预测麻疹的流行时间, 本文采用 ARIMA 模型拟合麻疹发病的资料, 同时探讨模型的可行性, 为麻疹防控工作提供依据。

1 资料与方法

1.1 资料 2009–2015 年乌鲁木齐市麻疹疫情资料来源于“国家疾病监测信息报告管理系统”和“乌鲁木齐市麻疹疫情专报系统”, 利用 2009 年 1 月–2015 年 12 月的麻疹月发病例数(84 个数据)建立预测模型, 并利

作者简介: 梁静(1989–), 女, 硕士, 研究方向: 疫苗相应传染病的监测、预测及决策。

用模型对乌鲁木齐市 2016 年的发病情况进行预测。

1.2 研究方法 ARIMA 模型全称为自回归滑动平均混合模型(autoregressive integrated moving average model, ARIMA), 是一种时间序列预测方法。按照模型是否包含季节性可以分为非季节性 $ARIMA(p, d, q)$ 模型、季节性 $ARIMA(P, D, Q)_s$ 模型和 $ARIMA(p, d, q)(P, D, Q)_s$ 模型。 p, d, q 和 P, D, Q 分别是非季节和季节性自回归(AR)、差分(I)、移动平均(MA)的阶, 是一个季节性周期所包含的时点数。ARIMA 建模法分为 4 个阶段: 序列的平稳化、模型识别、参数估计和模型诊断, 然后选择最优模型进行预测^[3–9]。

1.2.1 序列的平稳化 使原序列满足 ARIMA 模型平稳可逆的要求。

1.2.2 模型识别 主要通过读 ACF、PACF 把握模型的大致方向, 为目标序列定阶, 提供几个粗模型以便进一步分析完善。

1.2.3 参数估计和模型诊断 采用 Box–Jenkins 方法模型, 观察数据的自相关、偏相关函数, 如果数据自相关、偏相关函数既不截尾也不拖尾, 且不是呈线性衰减趋势, 就可以判断数据序列适合于 ARIMA 模型。诊断

检验模型是否合适,需要对其拟合优度进行检验,典型方法是对观测值和模拟值的残差进行分析。如果残差序列不是白噪声序列,说明模型中还有信息包含在相关的残差序列中未被提取,则需重新建立模型,重复上述步骤,直到残差序列是白噪声序列为止。

1.2.4 预测 根据模型进行预测。

1.3 统计分析 应用 SPSS17.0 对麻疹发病数资料拟合建立 ARIMA 模型。

2 结果

2.1 发病趋势分析及预处理(平稳化) 乌鲁木齐市 2009-2015 年各月的麻疹发病例数见表 1,依据其绘制时间序列图见图 1。图 1 可见,2009-2015 年麻疹发病例数呈现出周期性变化,每年 5-7 月出现发病高峰,存在着明显的季节波动性。对本序列进行平稳化处理,经过一阶季节差分($sd=1$)消除季节的影响,得到的新数据基本稳定,见图 2。

表 1 2009-2015 年乌鲁木齐市麻疹发病例数

年度	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月	合计
2009	1	8	9	22	31	28	10	2	5	4	4	6	130
2010	2	2	8	8	11	16	8	1	1	4	3	2	66
2011	4	13	19	47	79	77	14	5	6	2	0	6	272
2012	0	10	15	28	41	42	11	10	4	1	1	4	167
2013	5	6	14	34	126	75	26	2	2	0	0	1	291
2014	3	5	26	29	23	19	1	8	5	3	1	1	124
2015	5	11	11	12	26	16	2	5	11	9	18	28	154

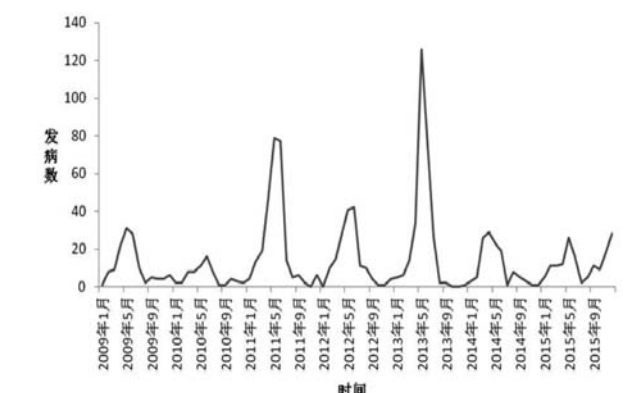


图 1 2009-2015 年乌鲁木齐市麻疹发病例数序列图

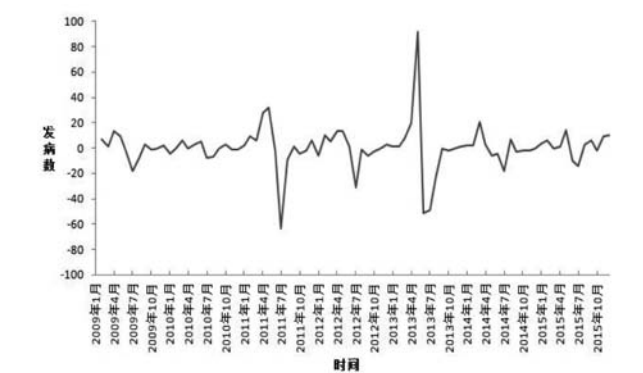


图 2 2009-2015 年乌鲁木齐市麻疹发病例数
平稳化处理($sd=1$)后的序列图

2.2 模型识别 经过一阶季节差分后,数据的季节性基本消除,故 $D=1$ 。且 ACF 图呈现逐渐衰减趋势(图 3、图 4)。可认为处理后的序列符合运用 ARIMA 模型的条件。初步确定模型为 $ARIMA(p,d,q)(P,1,Q)_{12}$,

这是一个综合了连续模型和季节模型的复合季节模型,其中季节以 12 个月为周期。

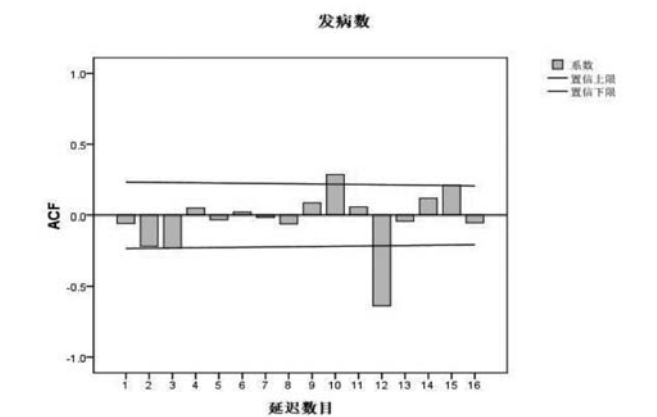


图 3 2009-2015 年各月乌鲁木齐市
麻疹发病数平稳处理($sd=1$)后的 ACF 图

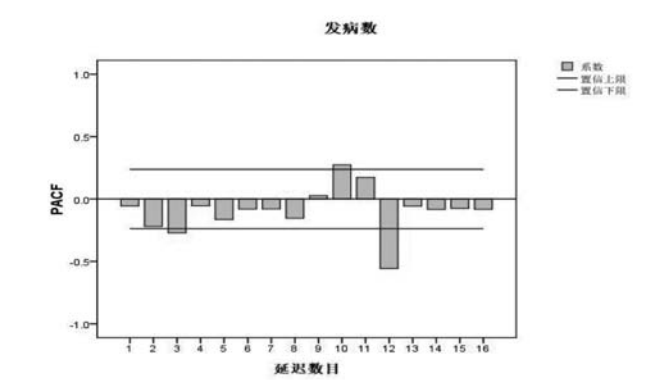


图 4 2009-2015 年各月乌鲁木齐市麻疹发病数
平稳化处理($sd=1$)后的 PACF 图

2.3 模型的建立 因模型初步确定为 $ARIMA(p,d,q)(P,1,Q)_{12}$ 模型,对于 p,d,q,P,Q 的识别,通过对参数采取从低阶到高阶逐个进行尝试的办法,而模型阶数过高会造成过拟合,故各阶数限定在 $0\sim 2$ 范围之内,对 $ARIMA(p,d,q)(P,1,Q)_{12}$ 模型进行参数估计,如令 p,d,q,P,Q 分别等于 0 和 1 时的各种不同组合进行拟合;然后通过 AIC 或 BIC 准则,从中选择最优模型,通过比较 $ARIMA(0,0,1)(1,1,0)_{12}$ 拟合效果最好,其中 $BIC=5.614$,滑动平均系数 $MA(1)=-0.481(t=4.531, P=0.000)$,季节回归系数 $AR(1)=-0.557(t=5.772, P=0.000)$,模型参数有统计学意义,见表 2。

表 2 2009–2015 年乌鲁木齐市麻疹月发病例数预测模型相关参数

参数	Estimate	SE	t	Sig.
MA(1)	-0.481	0.106	-4.531	0.000
AR(1), 季节性	-0.557	0.097	-5.772	0.000
常数	8.738	4.997	1.748	0.085

2.4 模型检验 利用第二步所建模型,得到实际值与预测值之差即残差,计算残差相关系数。对残差进行自相关和偏自相关分析显示,自相关系数均未超出 ± 0.5 ,因此不能认为残差序列各数值间有相关性(图 5),模型拟合较好;采用确定系数评价拟合效果,确定系数 $R^2=0.564, Q=16.086, P=0.447>0.05$,可认为残差数列是白噪声,模型拟合较好,说明建立的模型 $ARIMA(0,0,1)(1,1,0)_{12}$ 是合适的。

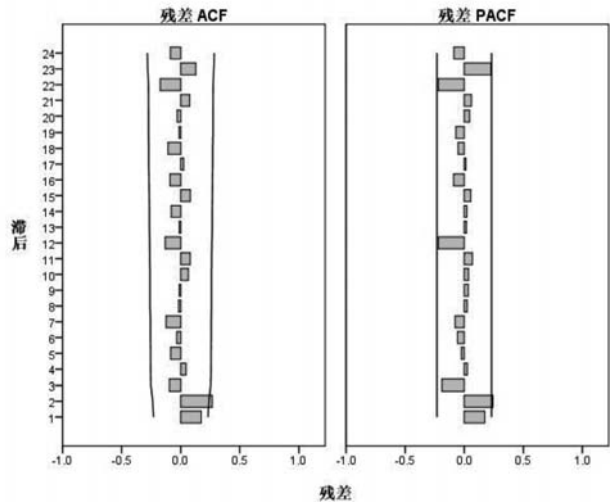


图 5 残差的自相关(ACF)和偏自相关(PACF)函数图

2.5 预测 利用建立的模型预测 2015 年乌鲁木齐市麻疹发病数,将麻疹拟合数据与 2015 年实际值相比,进行配对 T 检验, $t=5.79, P=0.574>0.05$,实际值与预

测值差异无统计学意义,可以看出(见表 3), $ARIMA(0,0,1)(1,1,0)_{12}$ 对麻疹发病数进行了较好的拟合,预测效果较好。现预测 2016 年各月麻疹发病例数即置信区间见表 4 和图 6。

表 3 2015 年乌鲁木齐市麻疹发病例数的实际值和模型预测值

月份	实际值	预测值	绝对误差	相对误差
1	5	4	-1	-0.20
2	11	6	-5	-0.45
3	11	22	11	1.00
4	12	27	15	1.25
5	26	73	47	1.81
6	16	27	9	0.56
7	2	9	7	3.50
8	5	1	-4	-0.80
9	11	5	-6	-0.55
10	9	4	-5	-0.56
11	18	3	-16	-0.89
12	28	8	-20	-0.71

表 4 乌鲁木齐市 2016 年麻疹月发病例数预测值与 95%置信区间

模型	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
预测	13	8	19	21	24	18	1	7	8	6	9	13
UCL	44	42	54	56	59	52	36	41	42	40	43	47
LCL	-18	-27	-15	-13	-10	-17	-33	-27	-27	-29	-26	-21

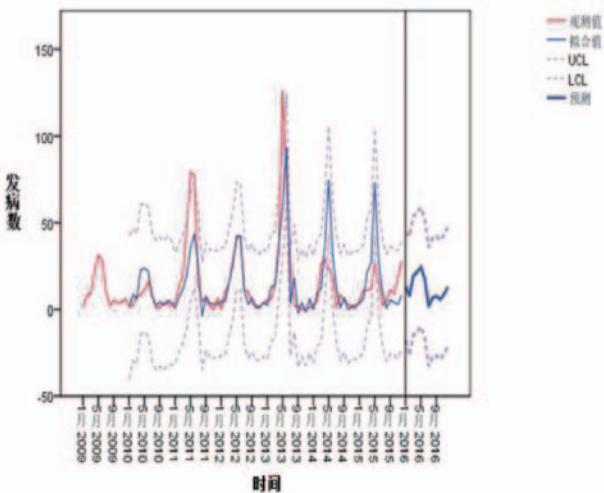


图 6 2016 年乌鲁木齐市麻疹发病例数与预测月发病例数曲线图

3 讨论

麻疹发病情况的预测研究是一项意义深远的工作^[8]。时间序列预测方法很多,但在实际研究中,有关影响因素数据资料无法获取时,时间序列大多非平稳,表现出各种趋势性和季节波动性,而 ARIMA 模型法是多个模型的混合,可以较好地预测短期数据。

ARIMA 模型根据序列识别一个试用模型,再加以诊断,做出必要调整,反复进行识别、估计、诊断的过程,直到建立适合的模型。ARIMA 可通过差分等方法将非平稳序列进行预处理,以满足预测的前提。ARIMA 使残差进入模型,更提高了模型的精度。

对乌鲁木齐市 2009–2015 年逐月麻疹发病例数数据分析发现乌鲁木齐市麻疹发病存在明显的季节性,5–7 月高发,每隔一年麻疹发病例数上升,然后来年降低,这为麻疹季节性和周期性防制提供了依据。

由于麻疹发病的影响因素很多,如人群易感性、免疫情况、卫生条件、环境、流动人口、发生疫情后采取的控制措施、其它地区发病情况对乌鲁木齐市的影响等,这些因素都会对预测效果产生影响。本研究证实了 ARIMA 乘积季节模型在麻疹发病的预测方面起到了一定的作用,但也应该注意到,由于疫情波动受到诸多未知随机因素的影响,所建立的模型不是一成不变的,它

较适合进行短期的预测,同时需要不断加入新的实际数据,以不断新拟合更能反映实际情况的预测模型,并提高预测的敏感性。

参考文献

- [1] 马超,罗会明,李黎,等. 世界卫生组织各区域和部分国家的消除麻疹免疫策略[J]. 中国疫苗和免疫,2014,20(3):254–263.
- [2] 国务院. 卫生事业发展“十二五”规划[EB/OL]. (2012–10–08)[2016–1–10]. http://www.gov.cn/zw/gk/2012-10/19/content_2246908.htm.
- [3] 刘刚,单芙蓉. ARIMA 模型及其在麻疹发病率预测中的应用[J]. 数理医药学杂志,2011,24(4):379–382.
- [4] Box, G, Jenkins G, Reinsel G. 时间序列分析预测与控制[M]. 顾岚,译. 第 1 版.北京:中国统计出版社,1997.
- [5] 方积乾,陆盈,张晋昕,等. 现代医学统计学(时间序列分析方法及其医学应用)[M]. 第 1 版.北京:人民卫生出版社,2002:219–269.
- [6] 孙振球. 医学统计学[M]. 第 1 版.北京:人民卫生出版社,2002:359.
- [7] 张文彤. SPSS 统计分析教程[M]. 第 1 版.北京:北京希望电子出版社,2002:557–582.
- [8] 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应用研究[J]. 中国卫生统计,2010,27(5):456–463.
- [9] 周权,聂俊雄,袁茂祥,等. ARIMA 模型在常德市痢疾发病预测中的应用[J]. 实用预防医学,2014,24(12):1454–1457.

收稿日期:2016–01–21

(接 1002 页)

结果显示:5 株副溶血性弧菌均携带 *tdh*、*tlh* 和 *toxR* 基因,不携带 *trh* 基因,通过结果快速地判定此次中毒事件的致病菌株为 VP,同时了解四种毒力基因携带情况。

综上所述,PathoMPS™ 系统检测技术作为一种检测病原微生物的方法,具有快速、方便、自动化程度高等优越性,这将在今后致病菌的检测中具有广阔的应用前景。可在疾病控制机构查找食物中毒致病菌中推广和运用,为事件处置和临床救治提供快速、及时、准确实验室依据。

参考文献

- [1] 侯瑞生,王丽,张勤,等. 多重 PCR 技术及其在病原体检测中的应用[J]. 中华全科医学,2011,9(8):1288–1290.
- [2] 林雪,陈泽辉,翁琴云. 应用多重实时荧光 PCR 方法检测副溶血性弧菌及其毒力基因[J]. 中国卫生检验杂志,2015,25(6):870–872.
- [3] Linda NW, Asim KB. Detection of *Vibrio parahaemolyticus* in shellfish by use of multiplexed real-time PCR with TaqMan fluorescent probes[J]. Appl Environ Microbiol, 2006, 72(3):2031–2042.
- [4] 李宏,杨大伟,刘云国,等. 多重荧光定量 PCR 同时检测霍乱弧菌、

副溶血性弧菌和创伤弧菌的方法研究[J]. 中国卫生检验杂志, 2011, 21(5):1180–1182.

- [5] 岳友宏,苏良. 双重 PCR 检测副溶血性弧菌的 *tlh* 和 *tdh* 基因方法建立与初步应用[J]. 实用预防医学, 2012, 19(9):1413–1415.
- [6] 徐苗苗,刘静雯. 副溶血性弧菌 *Vibrio parahaemolyticus* O3:K6 大流行克隆的溯源[J]. 微生物学通报,2014,41(10):2112–2121.
- [7] 程晓艳. 几种食源性致病菌快速检测技术的建立[D]. 青岛:中国海洋大学,2012.
- [8] 万志刚,汤慕瑾,吕敬章,等. 多种食源性致病菌检测的多重 PCR 方法的研究[J]. 现代生物医学进展,2012,12(11):2177–2181.
- [9] 张晓焱,王梅,周映霞. 多重 PCR 技术检测肉品中致病菌的应用研究[D]. 重庆:西南大学食品学院,2009.
- [10] 王琪,滕勇勇,何仕雯,等. 多重 PCR 快速检测 5 种重要致病性弧菌[J]. 中国卫生检验杂志,2014,24(24):3497–3500.
- [11] 肖勇,沙丹,凌霞. PCR 和毛细管电泳技术检测沙门氏菌等 5 种病原菌的方法研究[J]. 现代预防医学,2012,14(39):3611–3614.
- [12] 李婧,邱少富,刘雪林,等. 副溶血弧菌分子分型技术研究进展及应用[J]. 中国公共卫生,2012,28(11):1533–1536.
- [13] 韩海红,李宁,郭云昌. 副溶血性弧菌分子标志基因研究概况[J]. 微生物学报,2015,55(1):12–21.

收稿日期:2016–02–03