

# 采用自回归移动平均模型预测中国流感病例数

谭恩丽<sup>1</sup>, 侯慧玉<sup>2</sup>, 包海荣<sup>1</sup>, 滕雪娇<sup>3</sup>, 张顺先<sup>3</sup>, 李保娣<sup>4</sup>, 黄晓霞<sup>5\*</sup>

(1. 兰州大学 第一医院 老年呼吸科, 兰州 730000; 2. 河北省人民医院, 石家庄 050051;

3. 中国疾病预防控制中心 寄生虫病预防控制所, 上海 200025;

4. 甘肃省疾病预防控制中心, 兰州 730000; 5. 中国疾病预防控制中心 病毒病预防控制所, 北京 102206)

**摘要:**采用自回归移动平均模型 (ARIMA) 对中国 (不含中国港澳台地区) 流感月报告病例数进行预测研究, 为中国流行性感冒 (流感) 的预防控制提供参考依据。使用 SPSS 24.0 软件, 以 2006 年 1 月至 2016 年 12 月中国流感月报告病例数建立时间序列模型, 并以 2017 年 1~5 月的月报告病例数作为验证数据, 评估和筛选最优模型。以 2006 年 1 月至 2016 年 12 月中国流感月报告病例数为基础数据, 建立的最优模型为 ARIMA (4, 0, 4), 其平稳  $R^2 = 0.672$ , 标化 BIC = 18.032, Ljung-Box  $Q = 16.381$ ,  $P = 0.089$ 。对 2017 年 1~5 月的数据进行预测, 预测相对误差的平均值仅为 -3.25%。ARIMA 模型在预测中国流感月报告病例数方面效果较好, 但模型的建立和预测应用是个动态过程, 需不断根据积累的数据进行调整, 从而提高预测精度。

**关键词:**流行性感冒 (流感); 月报告病例数; 自回归移动平均模型

中图分类号: R373.1 R183.9 文献标识码: A 文章编号: 1000-8721(2017)05-0699-07

DOI: 10.13242/j.cnki.bingduxuebao.003221

流感是由流感病毒引起的一种急性呼吸道传染病。流感病毒根据病毒基质蛋白及核蛋白的抗原性及其基因特性的不同分为甲、乙、丙 (或 A、B、C) 3 型。甲型流感病毒根据分布在其表面的血凝素 (Hemagglutinin, HA) 和神经氨酸酶 (Neuraminidase, NA) 的蛋白结构和基因特性的不同又分为多个亚型, 理论上可组合  $H_xN_y$  多种亚型, 目前发现的 HA 和 NA 分别有 18 个 (H1~18) 和 11 个 (N1~11) 亚型。目前研究发现流感病毒存在种属感染的特性, 鸟类是流感病毒的天然宿主, 可以自然感染所有亚型的 A 型流感病毒<sup>[1]</sup>。感染人类的流感病毒包括 H1N1、H2N2、H3N2、甲型 H1N1 (Pandemic H1N1) 和 B 型流感。H1N1 亚型、H3N2 亚型和 B 型病毒在人群中流行表现为季节性和周期性, 故其引起的流感又称为季节性流感。近年发现的人类感染禽流感 H5N1、H9N2、H6N1、H7N9 和 H10N8 仅是偶然发生, 禽流感不能在人群中稳定存在并持续传播<sup>[2-5]</sup>。流感是一种严重危害全球和中国公众健康的呼吸道传染病<sup>[6]</sup>, 全球每年约有 5%~10% 的成人和 20%~30% 的儿童罹患流感, 300 万~500 万重症病例和 25 万~50 万死亡病例<sup>[7-9]</sup>。

近年来数理统计学中的数学模型 (如: ARIMA、神经网络模型等) 被应用于传染病疫情的预测和研判<sup>[10,11]</sup>, 能较好地揭示传染病随时间发展的动态变化, 可用于传染病的预测和预警。本文尝试建立中国流感月报告病例数的 ARIMA 模型, 为流感疫情的防控提供参考。

## 材料与方 法

### 1 资料来源

2006 年 1 月至 2017 年 5 月的中国流感月报告病例数 (不含中国港澳台地区) 来源于中国疾病预防控制中心主办的《疾病监测》杂志中每月公布的中国甲乙丙类传染病疫情动态简介。该数据由中国各级各类医疗卫生机构通过《传染病报告管理信息系统》报告。2006 年 1 月至 2016 年 12 月全国流感月报告病例数用来拟合 ARIMA 模型, 2017 年 1~5 月全国流感月报告病例数用于模型预测效果评价。

### 2 ARIMA 模型的建模方法

ARIMA 模型的建立包括以下步骤: ① 时间序列的特征分析及平稳化: 建立时间序列, 然后通过自相关系数函数图 (Autocorrelation function, ACF) 和偏相关系数函数图 (Partial autocorrelation function, PACF) 分析序列的平稳性, 如果序列不平稳, 将其通过自然对数转换、一般差分和/或季节差分等方法预处理数据。② 模型识别: 根据自相关分析图、偏自相关图和 SPSS 24.0 软件初步拟合结果,

收稿日期: 2017-07-21; 修回日期: 2017-08-25

作者简介: 谭恩丽 (1982-), 主治医师, 从事呼吸道疾病诊断和治疗工作, E-mail: 13659421557@163.com

\* 通讯作者: 黄晓霞 (1980-), 副研究员, 从事传染病的预防控制研究, E-mail: huangxiaoxia2012@163.com

尝试对模型进行初步识别和定阶。预先考虑使用 ARIMA(p, d, q) 或 ARIMA(p, d, q) × (P, D, Q) 模型, 其中 P、Q 一般不会超过 2 阶, 可依次取 0、1 和 2 并由低阶到高阶逐个尝试, 结合以往经验, p 和 q 一般都不会超过 5。③ 备选模型的参数估计和诊断: 通过非线性最小二乘法估计参数, 然后对模型的残差进行 Ljung-Box Q (白噪声) 检验, 以判断模型的合理性。依据模型是否简洁 (模型越简化越好)、ARIMA 各参数是否有统计学意义 ( $P < 0.05$ )、模型检验指标 (Box-Ljung 统计量及  $P$  值,  $P > 0.05$ )、平稳决定系数  $R^2$  (越大越好)、标准化贝叶斯准则 (BIC) (越小越好) 和均方根误差 (越小越好) 选择备选模型。④ 最优模型的确定: 在满足模型成立的基本条件之后, 选择备选模型, 用备选模型预测 2017 年 1~5 月中国流感报告病例数, 并与实际报告病例数进行对比, 选择预测相对误差最小的模型为最优模型。⑤ 模型的预测应用: 使用最优模型预测 2017 年 1~12 月中国月报告流感病例数。

### 3 统计学方法

采用 Excel 2010 软件对 2006 年 1 月至 2017 年 5 月的中国流感月报告病例数进行整理, 采用 SPSS 24.0 软件进行 ARIMA 模型预测分析, 通过比较预测相对误差的大小确定最优模型并进行预测应用。

## 结 果

### 1 中国流感报告病例发病趋势分析

根据 2006 年 1 月至 2016 年 12 月中国流感月报告病例数随时间的变化趋势 (图 1), 初步判断中国流感月报告病例的时间序列是非平稳曲线, 2006 年 8 月报告病例数最少 (417 例), 2016 年 3 月报告病例最多 (76 498 例)。2009 年 8 月~2010 年 1 月报告病例数较高, 且呈明显的双峰分布, 2016 年 1~4 月和 10~12 月份的报告病例数也呈双峰分布, 且明显高于其余年份。除 2009~2010 年和 2016 年出现两次高峰外, 中国流感病例呈现出一定的季节分布特征, 每年 11 月到次年 2~3 月为一个流行高峰; 2014~2015 年的 6~7 月还出现一个小高峰, 其他年份 6~7 月未出现流行小高峰。

### 2 序列平稳化及模型识别

对原始数据进行自然对数转化和 1 阶差分处理, 得到 ACF 图 (图 2) 和 PACF 图 (图 3)。如图 2

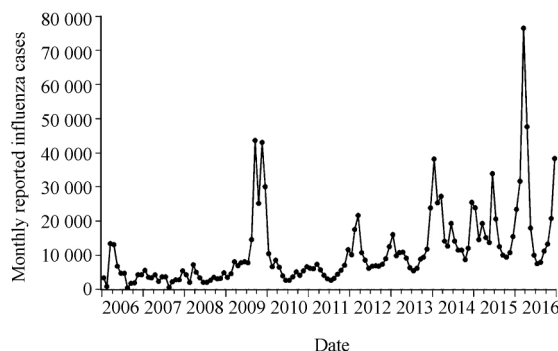


图 1 2006~2016 年全国流感月报告病例数时间序列图

Figure 1 Time series of monthly reported influenza cases in China, 2006-2016

和图 3 所示, 相关系数在延迟数目 12 时尚未落入 95% 的置信区间内, 初步说明需建立的最优 ARIMA(p, d, q) 或 ARIMA(p, d, q) × (P, D, Q) 模型将较为复杂, 即 p 和 q 会相对较高, 预设 p 和 q 不超过 5。通过从低到高逐步尝试的方法初步建立备选模型, 备选模型的拟合优度检验需要经过多次反复试验。

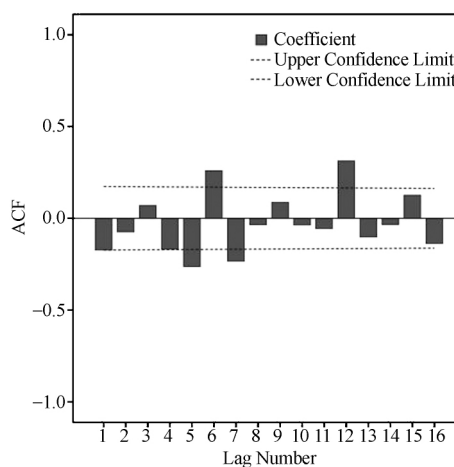


图 2 中国流感月报告病例数经自然对数转化和一阶差分后的 ACF 图

Figure 2 Time series after the natural-logarithm transformation and first-order difference of ACF for monthly reported influenza cases in China

### 3 备选模型的初步建立

在满足模型简洁、各参数都有统计学意义 ( $P < 0.05$ ) 和 Box-Ljung 统计量所对应的  $P > 0.05$  的情况下, 筛选出 12 个基本模型 (表 1)。从这些基本模型中选择平稳决定系数  $R^2$  最大 ARIMA(4, 0, 4) ( $0.672$ )、均方根误差最小 ARIMA(1, 0, 0) (1, 0,

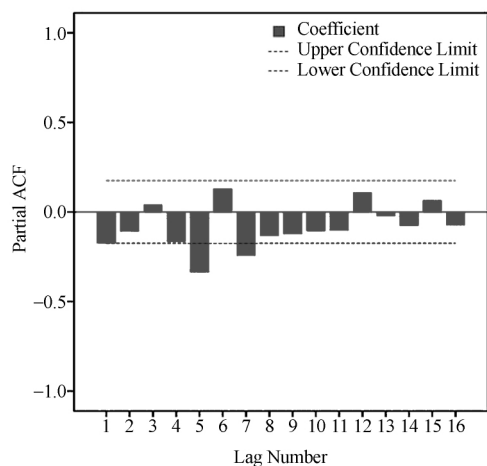


图3 中国流感月报告病例数经自然对数转化和一阶差分后的 PACF 图

Figure 3 Time series after the natural-logarithm transformation and first-order difference of PACF for monthly reported influenza cases in China

$(1,0,1)_{12}$  (6 940.523)、标化 BIC 最小 ARIMA(2,0,1)(1,0,1) $_{12}$  (17.927)、Ljung-Box Q 值最小 ARIMA(2,0,1)(1,0,1) $_{12}$  (14.026) 的 3 个模型作为备用模

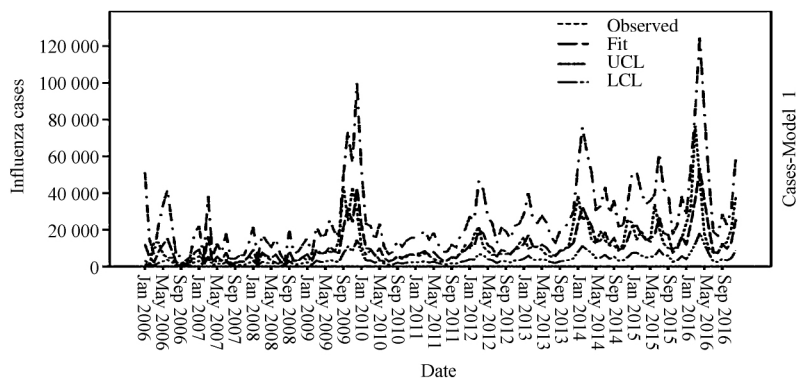


图4 ARIMA(4,0,4)模型拟合图

Figure 4 Fitting of an ARIMA(4, 0, 4) model

## 讨 论

近年来数理统计学在传染病领域的引入和应用,为传染病的防控提供了新的思路和方法,如 ARIMA 模型由于具备短期预测性强且简单易行等特点,已成为传染病领域中重要而基本的模型之一,广泛应用于传染病的预测和预警<sup>[10,12]</sup>。诸多自然因素影响流感疫情的发生和扩散,如人口流动、气温

和湿度等因素,收集这些变量资料的难度相对较高,

## 4 最优模型的确定

从表 1 可以看出在满足基本筛选条件的情况下,从不同角度出发,筛选的最优模型不完全一致。我们最终从预测值相对误差越小模型越优的原则出发,确定最优模型。分别以备用模型 ARIMA(1,0,0)(1,0,1) $_{12}$ 、ARIMA(4,0,4)和 ARIMA(2,0,1)(1,0,1) $_{12}$  预测 2017 年 1~5 月的中国流感月报告病例数(表 2),并与实际报告病例数进行比较,发现 ARIMA(4,0,4) 模型的预测相对误差最小(-3.25%),其中 2017 年 1 月和 2 月的预测值大于报告病例数,但 2017 年 3~5 月的预测值小于报告病例数(表 2)。其余 2 种模型的预测相对误差都很大(47.12%、52.82%),预测效果不理想。故最优模型为 ARIMA(4,0,4)(图 4)。

## 5 模型的预测应用

使用最优模型 ARIMA(4,0,4) 预测 2017 年 1~12 月中国月报告流感病例数,其中 2017 年 6 月为 20 566 例、7 月为 15 681 例、8 月为 14 361 例、9 月为 17 196 例、10 月为 13 226 例、11 月 12 811 例、12 月为 15 299 例(表 3)。

故使用回归分析预测我国流感月报告病例数一般难以实现,而将影响流感发生的社会、经济、气候和环境等因素的综合效应统一包括在时间这个唯一变量中,从而进行 ARIMA 预测,其短期预测效果比回归模型要好。

研究发现 2017 年 1~5 月流感的预测病例数和实际报告病例数之间相对误差较小(-3.25%),说明 ARIMA 模型在一定程度上可

准确估计我国的流感疫情,这与其他回归移动平均模型预测研究的结论相似<sup>[10]</sup>。另外,ARIMA模型与其它模型的联合使用可能会进一步提高我国流感病例数的预测准确性。有研究表明ARIMA和反向传播神经网络模型预测戊型肝炎病例数时误差为17.6%,但单一使用ARIMA进行预测时误差为25.0%<sup>[12]</sup>。在人工智能时代,

新算法(如卷积神经网络、深度置信网络等)的应用可使得预测结果更为准确。同时,尽可能收集影响流感等传染病的流行因素(如温度和湿度等气象条件、传播动力学参数等因素),并将这些因素作为自变量,也可在一定程度上提高ARIMA模型的预测精度。

表 1 根据 2006~2016 年中国流感月报告病例数建立 ARIMA 模型

Table 1 Establishment of ARIMA models according to monthly reported influenza cases in China, 2006~2016

模型 Model	参数 Parameter	估计值 Estimate	t	P	平稳 R <sup>2</sup> Stationary R-square	均方根误差 RMSE (Root-mean- square error)	标化 BIC Normalized BIC	Ljung -Box Q	df	P
ARIMA(1,0,0)(1,0,1) <sub>12</sub>	AR(1)	0.768	13.489	<0.001	0.65	6940.523	17.838	14.447	15	0.492
	SAR(1)	0.813	9.945	<0.001						
	SMA(1)	0.393	2.765	0.007						
ARIMA(1,0,0)(0,1,0) <sub>12</sub>	AR(1)	0.758	12.699	<0.001	0.579	9023.281	18.295	23.299	17	0.14
	AR(1)	0.738	11.9	<0.001						
ARIMA(1,0,0)(1,1,0) <sub>12</sub>	SAR(1)	-0.435	-4.94	<0.001	0.625	8355.008	18.157	14.628	16	0.552
	AR(1)	0.738	11.9	<0.001						
ARIMA(0,1,0)(1,1,0) <sub>12</sub>	SAR(1)	-0.425	-4.773	<0.001	0.094	10732.279	18.642	20.775	17	0.237
	AR(1)	0.707	10.847	<0.001						
ARIMA(1,0,0)(0,1,1) <sub>12</sub>	SMA(1)	0.53	5.776	<0.001	0.628	7751.041	18.031	21.494	16	0.16
	AR(1)	0.712	8.165	<0.001						
ARIMA(1,1,1)(1,0,0) <sub>12</sub>	MA(1)	0.997	2.865	0.005	0.196	7144.864	17.897	17.839	15	0.271
	SAR(1)	0.461	5.596	<0.001						
	SAR(1)	0.809	9.902	<0.001						
ARIMA(0,1,0)(1,0,1) <sub>12</sub>	SMA(1)	0.404	2.842	0.005	0.083	8243.883	18.146	22.768	16	0.12
	AR(1)	0.707	10.847	<0.001						
ARIMA(1,0,0)(0,1,1) <sub>12</sub>	SMA(1)	0.53	5.776	<0.001	0.628	7751.041	18.031	21.494	16	0.16
	SAR(1)	0.403	4.764	<0.001						
ARIMA(0,1,0)(2,0,0) <sub>12</sub>	SAR(2)	0.248	2.688	0.008	0.082	8438.11	18.193	22.618	16	0.124
	SMA(1)	-0.396	-4.613	<0.001						
ARIMA(0,1,0)(0,0,0) <sub>12</sub>	SMA(2)	-0.331	-3.322	0.001	0.093	8565.889	18.244	23.579	16	0.099
	AR(1)	-1.201	-17.6	<0.001						
ARIMA(4,0,4)	AR(2)	-0.363	-2.814	0.006	0.672	6971.672	18.032	16.381	10	0.089
	AR(3)	0.641	4.977	<0.001						
	AR(4)	0.789	12.318	<0.001						
	MA(1)	-2.104	-20.555	<0.001						
	MA(2)	-2.224	-10.803	<0.001						
	MA(3)	-1.309	-6.358	<0.001						
	MA(4)	-0.261	-2.538	0.012						
	AR(1)	1.719	11.909	<0.001						
ARIMA(2,0,1)(1,0,1) <sub>12</sub>	AR(2)	-0.724	-5.801	<0.001	0.651	6991.124	17.927	14.026	13	0.372
	MA(1)	0.968	8.191	<0.001						
	SAR(1)	0.816	9.647	<0.001						
	SMA(1)	0.422	2.857	0.005						

总之,模型的预测结果仅是一种参考,模型的建立和预测是个动态过程,需要不断根据积累的数据

进行调整,通过引入新算法来提高预测精度,可为流感等传染病的防控提供科学参考。

表 2 2017 年 1~5 月中国流感病例数预测值和实际值比较

Table 2 Comparison of predicted and actual values of influenza cases in China, January-May, 2017

月份 Month	报告病例数 Reported number	2017 年 1~5 月预测人数 Predicted influenza cases from January to May, 2017					
		ARIMA (1,0,0)		ARIMA (4,0,4)		ARIMA (2,0,1)	
		(1,0,1) <sub>12</sub>	相对误差(%) Relative error (%)	(1,0,1) <sub>12</sub>	相对误差(%) Relative error (%)	(1,0,1) <sub>12</sub>	相对误差(%) Relative error (%)
2017-Jan	30 109	41 422	35.57	30 589	1.59	42 025	39.58
2017-Feb	22 998	37 418	62.7	24 925	8.38	38 283	66.46
2017-Mar	30 519	52 077	70.64	28 212	-7.56	53 220	74.38
2017-Apr	23 260	35 836	54.07	20 316	-12.66	37 499	61.22
2017-May	19 085	21 496	12.63	17 936	-6.02	23 370	22.45
平均相对误差(中位数)							
Average relative error (median)	—	—	54.07	—	-6.02	—	61.22
平均相对误差(均数)							
Average relative error (mean)	—	—	47.12	—	-3.25	—	52.82

表 3 2017 年 1~12 月中国流感病例数预测

Table 3 Prediction of monthly number of influenza cases in China, January-December, 2017

预测值及 置信区间 Forecast and 95%CI	Jan-17	Feb-17	Mar-17	Apr-17	May-17	Jun-17	Jul-17	Aug-17	Sep-17	Oct-17	Nov-17	Dec-17
预测值 Forecast	30 589	24 925	28 212	20 316	17 936	20 566	15 681	14 361	17 194	13 226	12 811	15 299
95%上限 95%UCL	71 154	73 575	94 693	74 202	68 365	79 735	62 683	57 729	69 566	53 890	52 337	62 519
95%下线 95%LCL	10 381	5 496	4 745	2 806	2 227	2 444	1 718	1 550	1 824	1 376	1 323	1 579

## 参考文献:

- [1] 张焯, 李晓丹, 邹淑梅, 薄洪, 董丽波, 高荣保, 王大燕, 舒跃龙. 2009~2013 年我国活禽市场环境样本中禽流感病毒的检测[J]. 病毒学报, 2015, 31(6): 615-619. doi: 10.13242/j.cnki.bingduxuebao.002820.
- [2] 冯录召, 杨鹏, 张涛, 杨娟, 傅传喜, 秦颖, 张奕, 马春娜, 刘兆秋, 王全意, 赵根明, 余宏杰. 中国季节性流感疫苗应用技术指南(2014-2015)[J]. 中华流行病学杂志, 2014, 35(12): 1295-1319. doi: 10.3760/cma.j.issn.0254-6450.2014.12.001.
- [3] 李保娣, 李红育, 张慧, 于德山, 何健. 2013-2015 年度甘肃省流感病毒病原学特征分析[J]. 中华实验和临床病毒学杂志, 2016, 30(5): 448-451. doi: 10.3760/cma.j.issn.1003-9279.2016.05.008.
- [4] 杜宁, 王大燕, 舒跃龙. 2009 新甲型 H1N1 流感病毒病原学概述[J]. 病毒学报, 2009, 25(6): 479-484. doi: 10.13242/j.cnki.bingduxuebao.002042.
- [5] 李希妍, 成艳辉, 谭敏菊, 黄维娟, 郭俊峰, 隗合江, 肖宁, 蓝雨, 赵翔, 杨磊, 王钊, 王大燕, 舒跃龙. 2013~2014 年度中国 H3N2 亚型流感病毒病原学特征分析[J]. 病毒学报, 2015, 31(01): 30-35. doi: 10.13242/j.cnki.bingduxuebao.002622.
- [6] 夏瑜, 沈强, 陆步来, 雅雪蓉, 韩卫宁, 张钧, 吴伟华. 2010-2014 年苏州市流感病原学监测结果分析[J]. 中华实验和临床病毒学杂志, 2015, 29(6): 498-501. doi: 10.3760/cma.j.issn.1003-9279.2015.06.008.
- [7] World Health Organization. <http://www.who.int/immunization/topics/influenza/en/>.
- [8] World Health Organization. Vaccines against influenza WHO position paper—November 2012[J]. Wkly Epide-

- miol Rec. 2012;87(47):461-76.
- [9] 辛丽,鲁健,陈永坤,唐静,郭俊峰,王大燕,舒跃龙. 季节性流感疫苗株制备技术平台的建立[J]. 中华实验和临床病毒学杂志, 2015, 29(2):174-176. doi: 10.3760/cma.j.issn.1003-9279.2015.02.027.
- [10] 张顺先, 邱琪, 王英. 我国手足口病重症患者数自回归移动平均模型预测研究[J]. 病毒学报, 2017, 33(1): 77-81. doi: 10.13242/j.cnki.bingduxuebao.003097.
- [11] Wei W, Jiang J, Liang H, Gao L, Liang B, Huang J, Zang N, Liao Y, Yu J, Lai J, Qin F, Su J, Ye L, Chen H. Application of a combined model with autoregressive integrated moving average (arima) and generalized regression neural network (grnn) in forecasting hepatitis incidence in Heng County, China [J/OL]. PLoS One, 2016, 11(6):e0156768. doi: 10.1371/journal.pone.0156768.
- [12] Ren H, Li J, Yuan Z A, Hu J Y, Yu Y, Lu Y H. The development of a combined mathematical model to forecast the incidence of hepatitis E in Shanghai, China [J]. BMC Infect Dis, 2013, 13: 421. doi: 10.1186/1471-2334-13-421.

## Application of an Autoregressive Integrated Moving Average Model for the Prediction of Influenza Cases in China

TANG Enli<sup>1</sup>, HOU Huiyu<sup>2</sup>, BAO Hairong<sup>1</sup>, TENG Xuejiao<sup>3</sup>,  
ZHANG Shunxian<sup>3</sup>, LI Baodi<sup>4</sup>, HUNAG Xiaoxia<sup>5\*</sup>

(1. *Department of Gerontal Respiratory Medicine, the First Hospital of Lanzhou University, Lanzhou 730000, China;* 2. *Hebei General Hospital, Shijiazhuang 050051, China;*

3. *National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, Shanghai 200025, China;*

4. *Gansu Provincial Center for Disease Control and Prevention, Lanzhou 730000, China;*

5. *National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China)*

**Abstract:** An autoregressive integrated moving average (ARIMA) model was used to predict the monthly influenza cases in China to provide a reference for influenza prevention and control. SPSS v24.0 was used to construct the ARIMA model based on the monthly influenza cases from January 2006 to December 2016 year. The number of reported cases from January to May 2017 was used to evaluate and screen for the optimal model. The fitted model was ARIMA (4, 0, 4) based on cases from January 2006 to December 2016, stationary R-square=0.672, normalized BIC=18.032, Ljung-Box Q=16.381, P=0.089. The relationship prediction error was -3.25%. These data suggest that the establishment and prediction of an ARIMA model is a dynamic process that need to be adjusted continuously according to accumulated data.

**Key words:** Influenza; Monthly reported cases; ARIMA

---

\* Corresponding author: HUANG Xiaoxia, E-mail: huangxiaoxia2012@163.com