·专题·

ARIMA 季节模型在我国艾滋病发病预测中的应用

尤佳豪,张蓓蓓,丁 勇

(南京医科大学康达学院医学信息工程教研室,江苏 连云港 222000)

摘要:目的 探讨适合用于预测我国艾滋病月发病人数的模型,为艾滋病的预防提供参考。**方法** 收集 2006 年 1 月-2018 年 12 月全国艾滋病月发病数资料,建立 ARIMA 季节模型,对 2019 年 1 月-6 月艾滋病月发病数评估预测效果。**结果** 艾滋病月发病数呈明显季节性特征,ARIMA(1,1,2)(0,1,1)12 模型较好地拟合了既往艾滋病的实际发病序列,拟合优度 0.902,该模型的预测结果平均相对误差率为 10.10%,有较好的预测效果。**结论** ARIMA 模型能够较好地拟合并预测我国艾滋病的月发病人数,为艾滋病的防控提供定量分析的依据。

关键词:ARIMA 模型;艾滋病;疾病预测

中图分类号:R512.91 **文献标识码:**A **DOI:**10.3969/j.issn.1006-1959.2021.17.001

文章编号:1006-1959(2021)17-0001-03

Application of ARIMA Seasonal Model in Predicting the Incidence of AIDS in China

YOU Jia-hao,ZHANG Bei-bei,DING Yong

(Department of Medical Information Engineering,Kangda College,Nanjing Medical University,Lianyungang 222000,Jiangsu,China)

Abstract:Objective To explore a model suitable for predicting the monthly incidence of AIDS in China, and to provide references for the prevention of AIDS.**Methods** Data on the monthly incidence of AIDS in China from January 2006 to December 2018 were collected, and an ARIMA seasonal model was established to evaluate the prediction effect of the monthly incidence of AIDS from January to June 2019.**Results** The monthly incidence of AIDS showed obvious seasonal characteristics. The ARIMA(1,1,2)(0,1,1)12 model fitted the actual incidence sequence of previous AIDS well with a goodness of fit of 0.902.The average relative error rate of the prediction results of this model was 10.10%, which had a good prediction effect. **Conclusion** The ARIMA model can better fit and predict the monthly number of AIDS cases in China, and provide a quantitative analysis basis for AIDS prevention and control.

Key words:ARIMA model;AIDS;Disease prediction

AIDS 是一种危害性极大的传染病,由感染 HIV 引起。HIV 是一种能攻击人体免疫系统的病毒,将人体免疫系统中最重要的 $CD4^+T$ 淋巴细胞作为主要攻击目标,大量破坏该细胞,使人体丧失免疫功能,最后导致死亡。艾滋病主要通过性接触、血液接触、母婴传播等方式进行传播[1]。该病的防治是一项长期的重要任务,良好的预测能对未来近期艾滋病的预防和控制提供预警。国内外用于传染病预测的方法有很多,比较常用的有时间序列分析法[2]、动力学模型[3]、灰色预测等。随着计算机科学的应用和发展,预测理论借助计算机强大的计算能力也得到了较快的发展。预测理论分为 3 种,分别是定性预测、定量预测、综合预测。定性预测是通过对当地传染病的流行过程、流行特征及其有关因素的具体分析,判断该病即将流行的趋势和强度。定量预测是借助数学手段利用原始资料,建立恰当的数学模型,预测未来传染病的发病数和发病率。综合预测又称组合预测,是指应用 2 种或 2 种以上的预测模型对某种传染病进行预测,综合利用各种单个预测模型所提供的信息,以适当的加权平均形式得出组合预测模型。ARIMA 模型适用于各种复杂的时间序列模式,是目前较通用的预测方法之一[4-7],已广泛应用于传染病发病率的预测,特别是具有季节性趋势的传染病预测。本文收集我国艾滋病发病疫情数据,应用 ARIMA 模型拟合全国艾滋病的月发病率,并预测短期艾滋病发病趋势,旨在对这类传染病早期预警提供理论参考。

1 资料与方法

1.1 数据来源 数据资料来源于我国疾病预防控制局网站(http://www.nhc.gov.cn/jkj//new_index.shtml)2006 年 1 月-2019 年 6 月的全国法定报告传染病疫情资料,其中 2006 年 1 月-2018 年 12 月的数据用于建立模型,2019 年 1 月-6 月的数据用于验证模型的预测效果。

1.2 方法 建立季节性 ARIMA 模型,即 ARIMA(p,d,q)$(P,D,Q)_s$,其中 p、q 为自回归和移动平均阶数,P、Q 为季节性自回归和移动平均阶数,d、D 为非季节性和季节性差分次数,s 为季节周期。对数据进行数据平稳化处理,通过时序图初步判断序列是否平稳,若为不平稳序列,则针对序列不平稳的趋势性或周期性进行差分或季节性差分处理,实现序列的平稳化。①模型识别:对平稳序列做自相关图,根据自相关函数和偏自相关函数拖尾、截尾情况估计 p、d、q 值,建立备选模型;并根据贝叶斯准则(BIC)选择最优模型。②模型检验:选择残差检验的 Q 统计

基金项目:江苏省高校自然科学研究项目(编号:19KJD330001)

作者简介:尤佳豪(1997.10-),男,江苏无锡人,本科,主要从事医学信息分析和管理工作

通讯作者:丁勇(1956.8-),男,江苏淮安人,硕士,教授,主要从事生物统计工作

量检验，根据各滞后期 Q 统计量的 *P* 值，检验结果不能拒绝残差不相关的零假设，即模型的残差序列是白噪声序列，所选模型恰当，可用于预测。③预测并验证：运用最终选定的 ARIMA 模型进行预测，并与实际值比，计算残差的 95%*CI*（置信区间）以及相对误差，以验证模型的拟合效果。

1.3 统计学方法 采用 SPSS 23.0 软件进行数据统计分析，取显著性水平为 0.05。

2 结果

2.1 序列的平稳化 2006 年 1 月–2018 年 12 月我国艾滋病月发病数时间序列图见图 1，该序列呈现出明显的非平稳性和季节性（s=12），并随着时间呈现递增。数据经过对数转换、一阶差分和一阶季节差分后达到平稳，见图 2。

2.2 模型的识别与定阶 由于原始数据经过一阶差分和一阶季节差分后达到平稳，取 s=12，d=1，D=1；观察差分后的自相关图见图 3，ACF 滞后 1 阶后趋向 0，判断序列的自相关函数呈 1 阶截尾，故 p=1；观察差分后的偏相关图见图 4，PACF 滞后 2 阶后逐步趋向 0，判断序列的偏相关函数呈 2 阶拖尾，故 q=2。模型初步为 ARIMA$(1,1,2)(P,1,Q)_{12}$，季节模型的 P、Q 值较难判断，但根据文献，参数 P、Q 很少超过 2 阶，分别取 0、1、2（共有 9 个模型）由低阶到高阶摸索试验，结合模型的拟合优度、残差以及系数间的相关性进行估计，采用 Ljung–Box 方法检验残差白噪声，非白噪声模型排除。

2.3 参数估计及诊断 对 9 组模型进行检验，模型 ARIMA $(1,1,2)(0,1,1)_{12}$ 正态化 BIC 值（12.839）最小，R^2=0.902 最大，杨–博克斯统计量为 18.726，*P*=0.176，残差序列为白噪声；残差序列的自相关系数及偏相关系数均在 95%CI，见图 5，由此判断 ARIMA$(1,1,2)(0,1,1)_{12}$ 模型为最优模型。

2.4 模型预测 用 ARIMA $(1,1,2)(0,1,1)_{12}$ 模型预测全国 2019 年 1 月–6 月艾滋病月发病人数，并对实际数据进行预测精度的验证，见表 1，该模型的预测结果平均相对误差率为 10.10%，预测值比较接近实际值，该模型具有较好的预测功能。

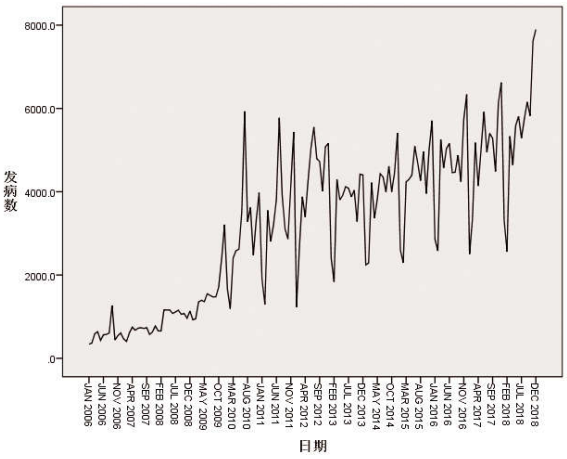

图 1 我国艾滋病月发病数时间序列图

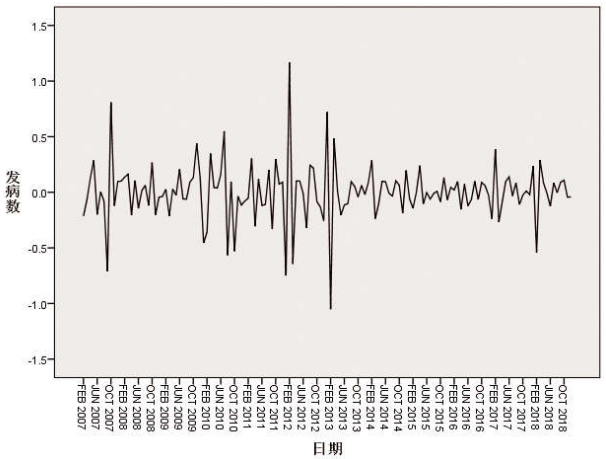

图 2 经过转换的数据序列图

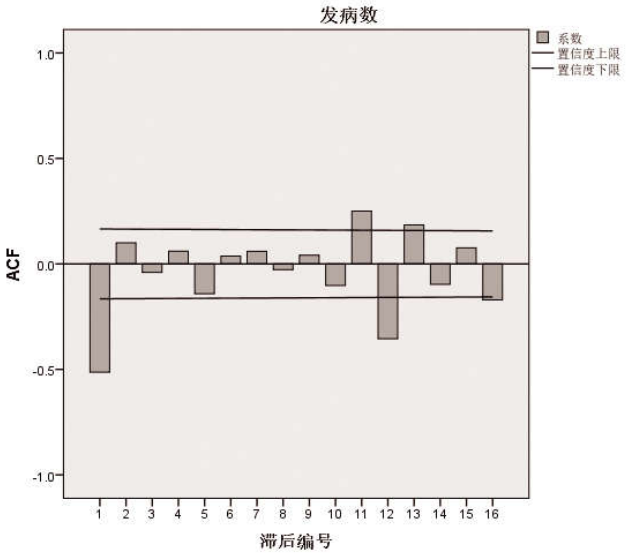

图 3 差分后序列的自相关图

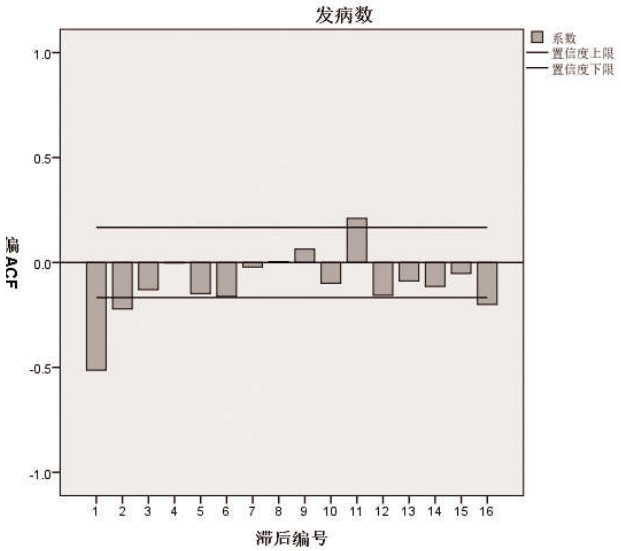

图 4 差分后序列的偏相关图

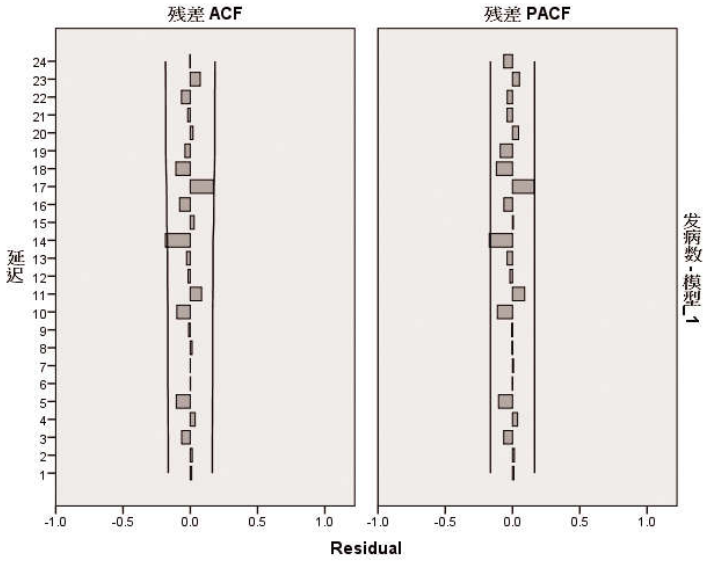

图 5 残差序列的自相关系数及偏相关系数

表 1 模型预测的误差

月份	实际值	估计值	相对误差
1	3688	4373.4	0.186
2	3587	4068.1	0.134
3	6086	6574.7	0.080
4	6277	5870.4	0.065
5	6291	6723.1	0.069
6	6642	7131.9	0.074

3 讨论

在定量预测模型中，ARIMA 模型能将各种已知的、未知的因素综合成统一的影响因素蕴含在时间序列变量中，比较灵活，既适用于非周期性序列，也适用于周期性序列。周期可以为年份、季度、月份，适用范围更广泛且所需的原始资料较少，对短期内传染病的预测效果较佳，具有较为广泛的应用前景。

目前国内对艾滋病的预测研究报告不多，且多是对艾滋病的地区年发病情况进行分析预测，对全国的发病情况进行预测研究的报道较少。本研究结果显示，全国 2006 年 1 月–2018 年 12 月艾滋病发病率呈现出明显季节周期性，且发病率呈逐年上升趋势，有必要对艾滋病发病趋势进行准确预测，提前做好应对措施、制定防控方案。本研究通过正态化 BIC 值最小，拟合优度最大，杨–博克斯统计量显著性和残差序列为白噪声等指标，筛选 ARIMA $(1,1,2)(0,1,1)_{12}$ 模型为拟合效果最优模型；同时利用 2019 年上半年艾滋病的月发病率进行预测，结果显示预测的平均误差绝对率为 10.10%，预测值接近真实值，提示该模型具有较好的预测功能。

建立 ARIMA 模型需要一定数量的历史数据，所建立的模型只能用于短期预测；当获得新数据时，应不断加入新的实际值，以修正或重新拟合更优的模型。因此，在制定艾滋病的预防控制策略和具体的措施时，还必须考虑其他综合因素对预测结果的影响，采用多种方法综合分析[8–10]，会有更好的效果和预测精度。

本文用 ARIMA 模型对我国艾滋病发病趋势进行了分析和预测，模型拟合优度为 0.902，预测结果的平均相对误差为 10.10%，说明 ARIMA 模型能够较好地拟合并预测我国艾滋病的月发病人数，为艾滋病的防控提供定量分析的依据。

参考文献：

[1]何纳.中国艾滋病流行新变化及新特征[J].上海预防医学,2019(12):1–6.

[2]张孟媛,张强,罗佳伟,等.重庆市艾滋病发病人数的 ARIMA 时间序列分析[J].中国卫生统计,2018,35(5):650–654.

[3]徐勇,张磊,凌莉.应用传染病动力学模型估计我国吸毒人群 HIV 年发病率[J].中华疾病控制杂志,2016,20(3):215–219.

[4]洪志敏,郝慧,房祥忠,等.ARIMA 模型在京津冀区域手足口病发病趋势预测中的应用[J].数理统计与管理,2018,37(2):191–197.

[5]刘芸男,彭荣荣,杨冬燕,等.ARIMA 模型在临床红细胞需求预测中的应用[J].安徽医科大学学报,2019,54(10):1611–1615.

[6]丁勇,吴静,武丹,等.ARIMA 乘积季节模型预测我国戊肝的发病趋势 [J]. 南京医科大学学报 (自然科学版),2020,40(11):1725–1729.

[7]包娅薇,邵明,陈雨婷,等.自回归求和滑动平均(ARIMA)模型在全球新型冠状病毒肺炎发病人数预测中的应用[J].中华疾病控制杂志,2020,24(5):543–548.

[8]汪鹏,彭颖,杨小兵.ARIMA 模型与 Holt–Winters 指数平滑模型在武汉市流感样病例预测中的应用 [J]. 现代预防医学,2018,45(3):385–389.

[9]李志超,刘升.基于 ARIMA 模型、灰色模型和回归模型的预测比较[J].统计与决策,2019(23):38–41.

[10]孙娜,许小珊,冯佳宁,等.ARIMA 与 GM(1,1)模型对我国肺结核年发病人数预测情况的比较[J].中国卫生统计,2019,36(1):71–74.

收稿日期：2021–04–11；修回日期：2021–04–25
编辑/肖婷婷