

ARIMA 乘积季节模型在湖南省 HIV 感染的应用研究

刘英¹, 唐玮¹, 赵天霄², 朱惠延^{1,2}

1. 南华大学公共卫生学院, 湖南 衡阳 421001; 2. 南华大学数理学院

摘要: **目的** 研究 ARIMA 乘积季节模型对湖南省 HIV 感染的适用性, 并进行预测。 **方法** 本文以湖南省 2005–2014 年 HIV 月感染数据建立模型, 以 2015 年月数据进行模拟。首先采用差分的方法对序列进行平稳化, 然后进行模型识别、定阶, 再对模型进行检验并预测未来 3 年的发病情况。 **结果** 对时间序列进行自然对数转换, 一阶差分和一阶季节差分得到 ARIMA 乘积季节模型, 结果显示 ARIMA(1, 1, 1)(0, 1, 1)₁₂ 很好地拟合了 HIV 的感染情况 ($R^2=0.894$)。残差序列经 Ljung-Box 检验为白噪声序列, $P=0.472$ 。预测结果显示未来 3 年全省 HIV 感染人数仍有增长趋势。 **结论** 利用乘积季节 ARIMA 模型对湖南省 HIV 感染拟合效果较好, 可以为卫生工作者采取控制措施提供依据。

关键词: 时间序列分析; 自回归滑动平均; HIV; 预测模型

中图分类号: R512.91 **文献标识码:** B **文章编号:** 1006-3110(2018)06-0760-04 **DOI:** 10.3969/j.issn.1006-3110.2018.06.035

艾滋病病毒(HIV)是一类人类免疫缺陷病毒, 主要破坏人类的免疫系统, 从而导致机体免疫功能缺失, 严重威胁着身体健康。我国自 1985 年首次报道, 感染人数逐年上升, 2000 年以后, 特别是 2005 年以来, 中国的艾滋病感染人数迅速上涨, 引起了人们的广泛关注。1992 年湖南省报告了首次病例, 全省感染人数也呈上升趋势。时间序列分析中的自回归滑动平均模型 (autoregressive integrated moving average model, ARIMA) 是由美国学者 Box 和英国统计学者 Jenkins 于 1976 年首次提出, 是由自回归模型和滑动平均模型组合而成^[1]。由于 HIV 感染数据是一列根据时间变化而呈现相应变化的序列, 具有趋势性、季节性、周期性和不规则性等, 是一类时间序列, 因此适用时间序列分析^[2]。有研究表明^[3], 时间序列分析在传染病发病率预测上具有应用价值。有研究者^[4-5]建立了乘积季节 ARIMA 模型预测了济宁市流行性腮腺炎和山东青岛市甲肝发病的情况。还有研究者^[6]利用时间序列模型预测甲肝的发病, 发现其具有较好的拟合效果。罗静等^[7]研究以重庆市艾滋病感染人群为例, 证明了 ARIMA 模型对艾滋病的新发感染率具有较好的拟合效果。本文拟根据湖南省近年 HIV 新发感染人数建立 ARIMA 乘积季节模型, 探究其在湖南省艾滋病疫情预测中的适用性。

1 资料与方法

基金项目: 湖南省教育厅科学研究重点项目(17A181)

作者简介: 刘英(1990–), 女, 湖南邵阳人, 硕士研究生, 研究方向: 流行病与卫生统计学。

通信作者: 朱惠延, E-mail: zhuhuiyan@126.com。

1.1 资料来源 2005–2015 年湖南省 HIV 感染数据来自湖南省疾病预防控制中心网络直报系统, 以新发感染人数作为样本。以 2005 年 1 月–2014 年 12 月湖南省月感染数据进行模型建立, 以 2015 年 1–12 月数据进行模型的拟合检验。

1.2 方法 ARIMA 方法是以时间序列的自相关分析为基础, 以识别时间序列的模式, 实现建模和完成预测的任务^[8]。模型的建立主要包括: 数据的预处理(平稳化); 模型的识别、定阶与模型参数估计; 模型的诊断检验^[9]。数据基本处理采用 Excel 2003, 模型建立采用 SPSS 19.0 软件, 对于序列平稳化采用 Eviews 8.0 进行单位根检验。

1.2.1 模型的平稳化 时间序列的平稳化是 ARIMA 模型分析的前提条件, 即要求均值不随时间变化, 方差不随时间变化; 自相关系数只与时间间隔有关, 与所处的时间无关。如果序列不平稳则需要通过差分使其平稳。在确定时间序列模型之前需把不平稳的时间序列转化为平稳序列。通常用下列方法^[10]: ①若序列呈线性趋势, 均值不平稳, 则采用一阶差分; ②若序列呈现二次趋势, 均值不为常数, 则采用二阶差分; ③若序列呈现随时间变化的趋势, 方差不为常数, 则通常采用自然对数使其平稳化。

1.2.2 模型的识别与定阶 根据序列的自相关图(ACF)和偏自相关图(PACF)看其是拖尾还是截尾, 初步为模型定阶。如果样本的自相关和偏自相关函数既不拖尾也不截尾, 不呈线性衰减趋势, 则不用 ARIMA 模型。相反地, 如果在周期点的整数倍上自相关或偏自相关函数呈现相当大的峰值并显示出振荡变化, 则可用 ARIMA 模型。然后根据图形是截尾还是拖

尾判断 p, q 值。

1.2.3 模型的诊断检验 模型是否合适,需要对其拟合度进行检验,典型方法是对观测值和模型拟合值的残差进行分析,即利用 SPSS 19.0 计算 Ljung-Box Q 统计量。如果残差序列不是白噪声序列,则说明还有信息包含在相关的残差序列中未被提取,模型其他参数不能完全代表建模对象的统计性质,即所建模型不是最终模型。

2 结果

2.1 数据的预处理 根据湖南省 2005 年 1 月-2014 年 12 月 HIV 感染人数的序列图。由图 1 可知,HIV 感染人数序列不是平稳序列,呈现一定的上升趋势。对数据进行自然对数转换和一阶差分,得到序列图,见图 2。

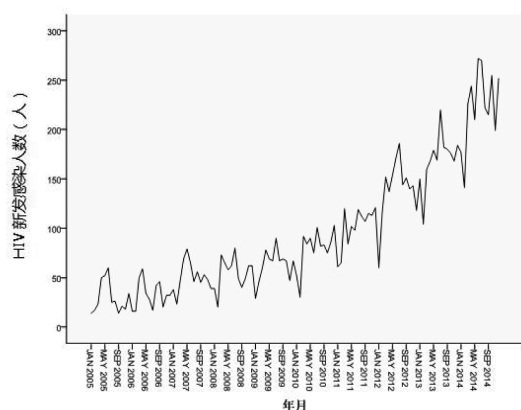


图 1 2005-2014 年各月感染数据序列图

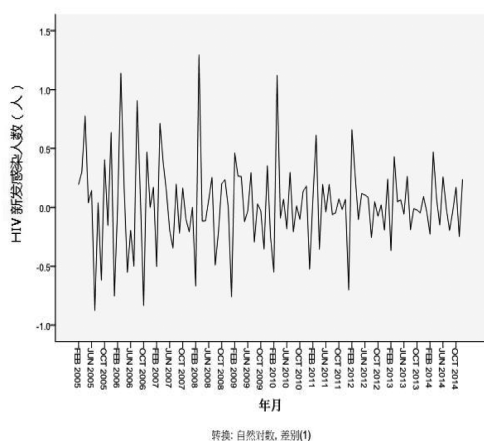
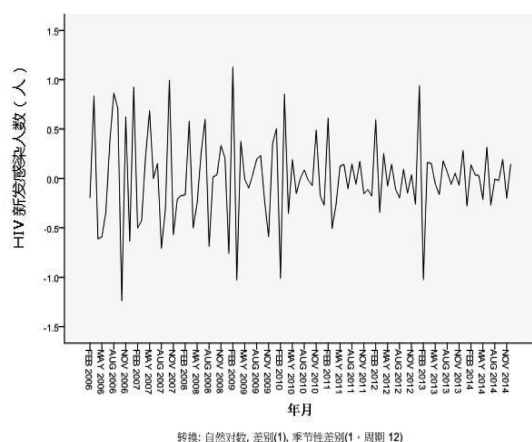


图 2 经对数转换和一阶差分的感染数据序列图

观察经对数变换和一阶差分后的时间序列自相关图,发现其存在周期性波动,因此,再进行季节性差分得到序列图,序列基本平稳,见图 3。利用 Eviews 8.0 进行单位根检验得到 $AIC = -0.3887, P = 0.0032$,序列平稳。

2.2 模型识别和定阶 根据上述分析,可初步尝试

建立 $ARIMA(p, d, q)(P, D, Q)_{12}$ 模型,即 ARIMA 乘积季节模型。由之前的分析可知 $d = D = 1$ 。



转换: 自然对数, 差(1), 季节性差(1·周期 12)

图 3 经自然对数转换,一阶差分和季节性差分序列图

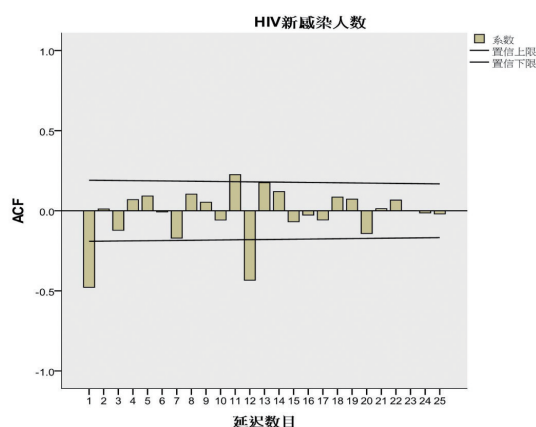


图 4 序列的自相关(函数)图

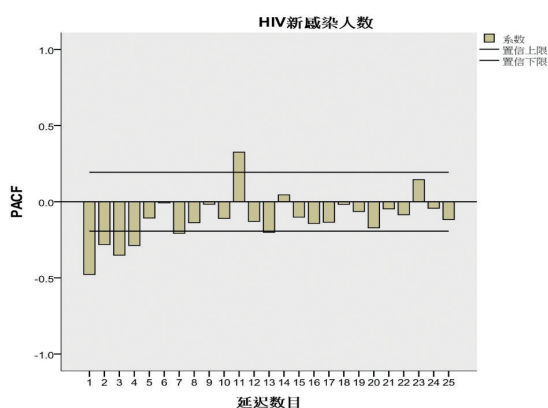


图 5 序列的偏自相关(函数)图

通过观察序列的自相关图和偏自相关图可以看出图形存在截尾性,自相关图可知,所有阶之后函数值除周期点处外基本都落在区间内,1 阶和 12 阶超出区间,所以 q 取 1, Q 取 1。通过偏自相关图,前 4 阶超出区间,12 阶未超出区间,所以 p 取 4, P 取 0。所以模型初定 $ARIMA(4, 1, 1)(0, 1, 1)_{12}$ 。见图 4、图 5。

2.3 建立模型 通过上述确定的模型 $ARIMA(4,1,1)(0,1,1)_{12}$ 带入软件 SPSS 19.0,得到参数估计结果。由于 $ARIMA(4,1,1)(0,1,1)_{12}$ 即 ARIMA1,其中自回归阶数 4 较大,模型较复杂,为此,把 4 改为 1 得到模

型 ARIMA2,即 $ARIMA(1,1,1)(0,1,1)_{12}$ 。带入 SPSS 19.0 得到模型的拟合结果,模型 ARIMA1 和 2 比较结果见表 1。

表 1 模型的拟合统计量

模型	模型拟合统计量							Ljung-Box Q	
	平稳的 R^2	R^2	RMSE	MAPE	MAE	MaxAPE	MaxAE	正态化的 BIC	统计量 Sig.
ARIMA1	0.569	0.894	21.655	22.501	16.136	125.469	55.313	6.500	11.630 0.476
ARIMA2	0.559	0.894	21.289	22.728	16.202	144.271	55.970	6.335	14.448 0.492

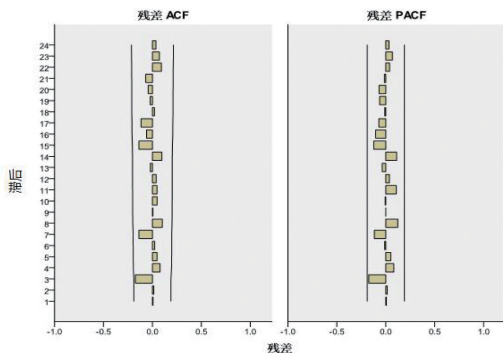


图 6 序列的残差自相关图和残差偏自相关图

由表 1 可知,模型 ARIMA1 对湖南省的 HIV 感染人数拟合效果较好, R^2 为 0.894,且序列的残差为白噪声序列,因此模型是可行的,见图 6。但取 4 阶模型较为复杂,改为 1 阶得到的 ARIMA2 模型,其对拟合度无太大影响,模型的决定系数 R^2 仍然为 0.894,误差项(均方根误差、平均绝对误差百分比、平均绝对误差、最大绝对误差百分比、最大绝对误差)略微有所增加。标准化的 BIC 减少,说明 ARIMA2 模型对数据的解释能力更大,且序列的残差仍然为白噪声($P>0.05$),因此采用更为简单的 $ARIMA(1,1,1)(0,1,1)_{12}$ 是可行的。

2.4 模型的预测 见表 2。2015 年 1-12 月份湖南省 HIV 感染的真实值与预测值稍有不同,但是都落在其 95%可信区间内,相对误差较低。由图 7 也可以看出 2005-2014 年感染的真实值在可信区间之内,由此说明模型对于湖南省 HIV 的预测是可行的。

表 2 利用 $ARIMA(1,1,1)(0,1,1)_{12}$ 对 2015 年 12 个月的预测

月份	预测值	真实值	相对误差	UCL	LCL
1	211	209	0.010	344	121
2	186	156	0.192	304	106
3	311	353	-0.119	508	178
4	311	296	0.051	507	177
5	305	279	0.093	498	174
6	321	263	0.221	524	183
7	353	316	0.117	576	202
8	305	279	0.093	498	174
9	301	275	0.095	492	172
10	321	282	0.138	524	183
11	310	255	0.216	506	177
12	362	271	0.336	591	207

注:UCL:预测值的 95%可信区间上限; LCL:预测值的 95%可信区间下限; 相对误差=(预测值-真实值)/真实值。

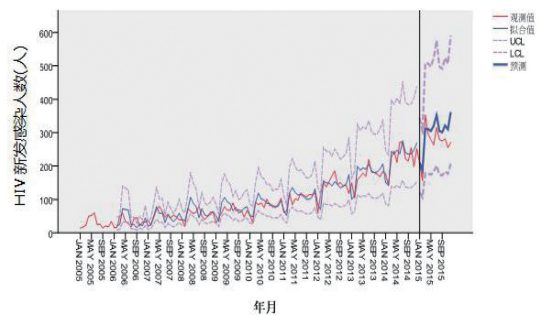


图 7 模型的拟合图

2.5 预测 根据 $ARIMA(1,1,1)(0,1,1)_{12}$ 模型预测 2016-2018 年湖南省各月的感染情况,结果见表 3。从表可知,各月的感染人数仍呈现上升趋势,这需要引起重视。

表 3 2016-2018 年预测值

年份		1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
2016	预测	282	246	408	404	396	416	458	397	396	426	416	492
	UCL	470	410	680	674	660	693	764	663	660	710	693	821
	LCL	157	136	226	225	220	231	255	221	220	236	231	273
2017	预测	379	326	537	529	516	542	599	522	523	567	561	674
	UCL	645	555	914	901	879	923	1 019	888	889	966	955	1 147
	LCL	205	176	290	286	279	293	324	282	282	307	303	364
2018	预测	512	436	711	698	678	712	787	689	695	761	762	928
	UCL	888	757	1 235	1 211	1 177	1 235	1 367	1 195	1 206	1 322	1 323	1 612
	LCL	269	230	374	367	357	375	414	362	366	401	401	489

3 讨 论

传染病的预测对于疾病控制具有重大意义^[11]:1、便于开展防治工作,有利于提高防治效果;2、根据预测的结果和实际结果进行比较,可以判断疾病是否按照以往的发展趋势,从而判断防治措施是否有效。如果真实值与预测值相当,在可信区间之内则说明疾病和之前的发展趋势一般无二;如果真实值低于预测值,则说明防治措施取得了较好的效果;如果真实值高于预测值则可能说明感染出现了不同于以往的流行趋势,可能出现新的危险因素,应警惕疾病暴发或流行的可能。

HIV 感染严重威胁着人类的健康,给个人、家庭和社会都会带来极大的负担,因此对 HIV 感染情况进行预测具有重大的意义。时间序列分析要求时间序列的残差是平稳的白噪声序列,如果序列残差不是白噪声序列则需要重新处理,重新建模。HIV 感染具有一定的趋势性,季节性和周期性,因此对于序列的平稳化是一个重要步骤。有研究者^[10]利用时间序列分析,建立了 ARIMA 季节乘积模型,并拟合肾综合症出血热,研究表明其具有良好的预测效果。本文根据湖南省 2005-2015 年各月的 HIV 感染数据进行时间序列分析,并建立了时间序列模型,采用了季节性的乘积模型,研究发现该模型拟合效果好,能较好的拟合湖南省各月的 HIV 感染情况。2015 年预测结果显示,除 3 月份之外,其他各月均存在一个高估的状态,可能随着人们对疾病的认识加深,医疗卫生机构对疾病的控制力加大,控制效果与往年相比存在优势;也可能是由于疾病监测系统更完善,报告更及时、完整;还有可能是因为随着人们对 HIV 认知的加强,具有危险行为的人自愿参与检查,从而发现更多病例。本文针对的是湖南省 2005-2015 年 HIV 感染人数进行预测,免去了计算率的误差,直接针对病例数更为直观。

根据建立的模型预测未来 3 年情况,为采取控制措施提供理论支持。预测结果显示湖南省未来三年的 HIV 新发感染状况仍表现为上升趋势,这需要相关部门采取相应措施加以控制。时间序列的研究不能存在缺失值(即感染人数为 0,建立模型对于湖南省大部分市不具有可行性,因为有些市月报告病例数为 0)。

控制 HIV 感染经性传播、血液传播、母婴传播对于

控制 HIV 新发感染具有重大作用,这些不仅仅是医疗卫生工作者的责任,也需要社会及个人共同参与,共同努力。由于现在 HIV 感染的传播方式以性传播为主^[12],且控制性传播又是最为方便和可控的,因此加强性安全的健康教育,鼓励群众安全性行为具有较为突出的作用。

建立时间序列模型受到很多因素的影响,因此建立的模型并非一成不变的,这是因为影响艾滋病的因素很多,主要是人类的活动和意识,一旦这些发生改变,从一定程度上会影响 HIV 感染状况,使得原有拟合模型不再适用或分析性能降低,所以并不能进行长期预测。另一方面本次研究主要是针对全省的感染总情况进行预测,由于各市的感染情况、感染主要方式都不尽相同,所以对于模型在各市的预测应用还需要进一步的研究。

参考文献

- [1] Geoge EP, Gwilym M. 时间序列分析预测与控制[M]. 北京:中国统计出版社,1997:12-13.
- [2] 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]. 北京:人民卫生出版社, 2002:219-269.
- [3] 冯超, 白杉. 时间序列模型拟合艾滋病发病趋势预测[J]. 中国公共卫生, 2005, 21(7):893.
- [4] 李润滋, 章涛, 梁玉民, 等. SARIMA 模型在流行性腮腺炎发病预测中的应用[J]. 山东大学学报(医学版), 2016, 54(9):82-86, 96.
- [5] 梁纪伟, 姜法春, 韩雅琳, 等. 应用 ARIMA 乘积季节模型预测青岛市甲肝病[J]. 中国公共卫生管理, 2016, 32(6):780-782, 793.
- [6] 刘继恒, 贺圆圆, 张皓, 等. 基于时间序列模型对甲型病毒性肝炎的预测研究[J]. 实用预防医学, 2017, 24(8):1009-1011.
- [7] 罗静, 杨书, 张强, 等. 时间序列 ARIMA 模型在艾滋病疫情预测中的应用[J]. 重庆医学, 2012, 41(13):1255-1256, 1259.
- [8] 孙振球, 徐勇勇. 医学统计学[M]. 第 4 版. 北京:人民卫生出版社, 2014:391.
- [9] Peter J. Brockwell and Richard A. et al 著, 田铮译. 时间序列的理论与方法[M]. 第 2 版. 北京:高等教育出版社, 2001:214.
- [10] 彭志行, 鲍昌俊, 赵杨, 等. ARIMA 乘积季节模型及其在传染病发病预测中的应用[J]. 数理统计与管理, 2008, 27(2):362-368.
- [11] 吴家兵, 叶临湘, 尤尔科. 时间序列模型在传染病发病率预测中的应用[J]. 中国卫生统计, 2006, 23(3):276.
- [12] 中国疾病预防控制中心性病艾滋病预防控制中心, 性病控制中心. 2017 年第 2 季度全国艾滋病性病疫情[J]. 中国艾滋病性病, 2017, 23(8):677.

收稿日期:2017-07-19