·方法研究·

ARIMA 预测模型在甘肃省其他感染性腹泻发病预测中的应用

刘希波[1],曹静[1],王云[1],李明阳[1],王淑霞[1],胡继宏[2]

1. 甘肃中医药大学公共卫生学院,甘肃 兰州 730000; 2. 甘肃中医药大学科研实验中心,甘肃 兰州 730000

摘要: 目的 建立并评价甘肃省其他感染性腹泻发病的 ARIMA 预测模型。 **方法** 利用 2010—2018 年甘肃省其他感染性腹泻的发病数据建立 ARIMA 预测模型,同时利用 2019 年发病数据评价模型并对 2020 年甘肃省其他感染性腹泻发病进行预测。 **结果** 根据模型拟合效果,模型 ARIMA(0,1,1)(1,1,0)$_{12}$为最优模型。R^2=0.741,Ljung-Box 检验值为 25.944,BIC 值为 11.060。模型拟合甘肃省其他感染性腹泻的发病趋势与实际发病趋势一致,MAPE=17.297%,预测结果显示 2020 年甘肃省其他感染性腹泻发病时间分布与往年趋于一致。 **结论** ARIMA(0,1,1)(1,1,0)$_{12}$模型能较好地拟合甘肃省其他感染性腹泻的发病趋势,对该病的预防控制、风险评估等具有一定的公共卫生意义。

关键词: ARIMA;预测模型;其他感染性腹泻

中图分类号:R181.2 **文献标识码:**A **文章编号:**1006-3110(2021)01-0113-04 DOI:10.3969/j.issn.1006-3110.2021.01.031

Application of ARIMA prediction model in forecasting the incidence of other infectious diarrhea in Gansu Province

LIU Xi-bo[1], CAO Jing[1], WANG Yun[1], LI Ming-yang[1], WANG Shu-xia[1], HU Ji-hong[2]

1.*School of Public Health, Gansu University of Chinese Medicine, Lanzhou, Gansu* 730000, *China*;

2.*Scientific Research and Experiment Center, Gansu University of Chinese Medicine, Lanzhou, Gansu* 730000, *China*

*Corresponding author: HU Ji-hong, E-mail:hujihonghappy@*163.*com*

Abstract: Objective To establish an autoregressive integrated moving average (ARIMA) prediction model in predicting the incidence of other infectious diarrhea in Gansu Province, and to evaluate its prediction effect. **Methods** Data regarding the incidence of other infectious diarrhea in Gansu Province from 2010 to 2018 were used to establish an ARIMA prediction model. At the same time, data about the incidence of other infectious diarrhea in 2019 were applied to evaluating the model, and the incidence of other infectious diarrhea in Gansu Province in 2020 was forecasted. **Results** The ARIMA (0,1,1)(1,1,0)$_{12}$ model was supposed to be the best fitted model. The value of R^2 was 0.741, the value of Ljung-Box Q statistic was 25.944, and BIC value was 11.060. The model fitted the incidence trend of other infectious diarrhea in Gansu Province, which was consistent with the actual incidence trend, and the mean absolute percentage error (MAPE) was 17.297%. The predicted results revealed that the time distribution of incidence of other infectious diarrhea in Gansu Province in 2020 tended to be similar to that of previous years. **Conclusions** The data predicted by the ARIMA(0,1,1)(1,1,0)$_{12}$ model can fit well with the incidence trend of other infectious diarrhea in Gansu Province, and the model has certain public health significance in prevention, control and risk assessment of other infectious diarrhea.

Keywords: autoregressive integrated moving average (ARIMA); prediction model; other infectious diarrhea

其他感染性腹泻(other infectious diarrhea)指除霍乱、细菌性和阿米巴性痢疾、伤寒和副伤寒以外的感染性腹泻病。自 2010 年以来,我国其他感染性腹泻的发病率一直位居全国法定报告的丙类传染病第 2 位[1],对各年龄人群尤其是婴幼儿健康造成极大威胁。其中 5 岁以下儿童其他感染性腹泻发病数占全国发病总数的 50%以上[2],是造成 5 岁以下儿童营养不良的主要原因。若儿童时期患有腹泻病,其沉重疾病负担会影响儿童生理和心理发育,并最终导致成年后的适应力和生产力损失[3]。2010—2016 年甘肃省其他感染性腹泻发病率逐年升高[4],2015 年,上升至甘肃省丙类传染病发病的首位[5]。目前甘肃省其他感染性腹泻流行病学相关研究报告尚未发现。本文利用 ARIMA 模型拟合甘肃省其他感染性腹泻发病的时间序列,建立并评估该模型在甘肃省其他感染性腹泻预测研究中

基金项目:甘肃中医药大学研究生创新基金(CX2019-46)

作者简介:刘希波(1993-),男,在读研究生,研究方向:公共卫生。

通信作者:胡继宏,E-mail:hujihonghappy@163.com。

的应用。

1 资料与方法

1.1 资料来源 甘肃省其他感染性腹泻 2010—2015 年报告病例数下载自 http://www.nhc.gov.cn/jkj/国家卫健委疾病预防控制局,2016—2018 年报告病例数下载自甘肃省卫健委官方网站(http://wsjk.gansu.gov.cn/)。

1.2 方法 利用 ARIMA(autoregressive integrated moving average)序列分析预测模型以拟合季节性的时间序列,即 ARIMA$(p, d, q)(P,D,Q)_{12}$模型[6]。d, D 是非季节性和季节性差分次数;p, q 是自回归阶数和移动平均阶数;P, Q 是季节性自回归阶数和移动平均阶数。建模步骤:(1)模型识别:依据时间序列平稳与否判断是否需要差分即差分次数;利用自相关函数(autocorrelation function, ACF)和偏自相关函数(partial autocorrelation function, PACF)分析时间序列的平稳性和季节性,初步确定模型。(2)模型估计:根据初步确定的模型,ARIMA 程序能够估计模型的参数。(3)模型诊断:利用决定系数(R^2)、Ljung-Box 检验、贝叶斯信息量(Bayesian information criterion,BIC)和平均绝对误差百分比(mean absolute percentage error, MAPE)进行模型诊断[6]。

1.3 统计学分析 采用 SPSS 22.0 进行数据分析,用 2010—2018 年甘肃省其他感染性腹泻报告病例数描述时间趋势及建立模型,利用 2019 年报告病例数评价模型拟合效果并预测 2020 年发病情况,检验水准 $\alpha=0.05$。

2 结 果

2.1 发病时间趋势 2010—2016 年,甘肃省其他感染性腹泻发病情况呈逐年增加趋势。季节性分布明显,下半年发病数明显高于上半年,发病高峰一般出现在 8 月和 11 月,见图 1。

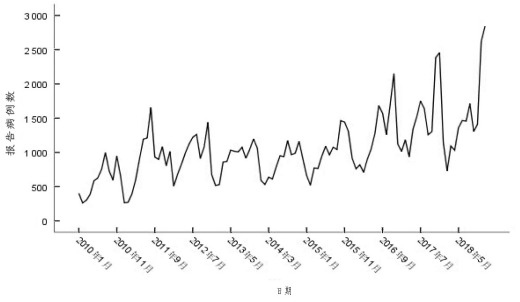
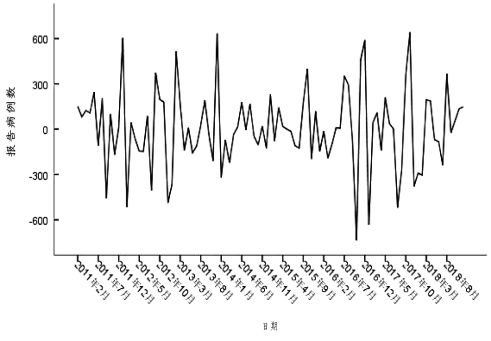


图 1 2010—2018 年甘肃省其他感染性腹泻报告病例数时间序列

2.2 模型识别 甘肃省其他感染性腹泻发病的时间序列(图 1)显示,存在明显的季节性和周期性。为使时间序列达到平稳性要求,对该序列进行一阶差分和一阶季节差分(图 2)。经平稳化后的时间序列的均值始终表现为在 0 的位置,基本满足平稳性要求。因此差分和季节差分均取 1,即 d 和 D 均为 1。得到 ACF 图(图 3)和 PACF 图(图 4),ACF 图和 PACF 图显示大部分阶数的自相关系数都在 2 倍标准差范围内波动,可以判定序列平稳。

ACF 图和 PACF 图显示,自相关系数在 2 阶截尾,偏自相关系数拖尾。选择滑动平均阶数 $q=1$,自回归阶数 $p=0$。季节性部分 P, Q 分别取 0,1,2,进行试验[7-8]。经过多次验证,备选模型为:ARIMA$(0,1,1)(0,1,1)_{12}$,ARIMA$(0,1,1)(1,1,0)_{12}$。



注:变换:求差(1),周期性差值(1,周期性 12)。

图 2 一阶差分和一阶季节性差分后甘肃省其他感染性腹泻报告病例数时间序列

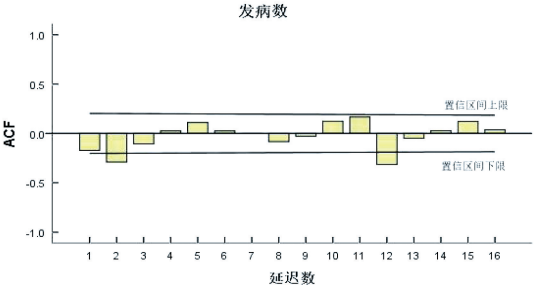


图 3 2010—2018 年甘肃省其他感染性腹泻报告病例数经平稳化后的 ACF 图

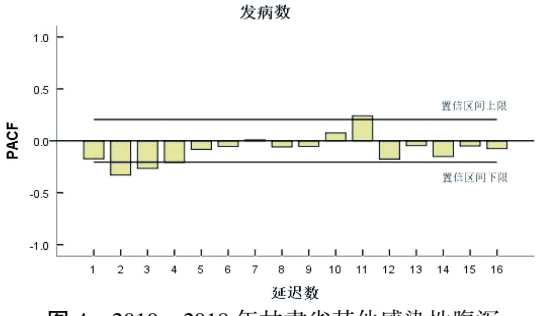


图 4 2010—2018 年甘肃省其他感染性腹泻报告病例数经平稳化后的 PACF 图

2.3 模型诊断 模型诊断建立在模型拟合优度

(R^2)、Ljung-Box 检验和 BIC 基础上[6,9]。Ljung-Box 检验模型残差是否为白噪声序列,否则排除模型,备选模型 P 值均大于 0.05,残差序列均为白噪声。BIC 值为 11.060 时最小,最优模型为 ARIMA(0,1,1)(1,1,0)$_{12}$(表 1),其 MAPE 为 17.297%,滑动平均(MA)滞后 1 阶的 P 值小于 0.001,自回归(AR)季节性滞后 1 阶的 P 值等于 0.001(表 2)。模型残差序列的自相关系数和偏自相关系数均在 95%置信区间内,且有趋于零的趋势,满足平稳性要求(图 5)。

表 1　模型比较

模型	R^2	Ljung-Box 检验	P 值	BIC 值
ARIMA(0,1,1)(0,1,1)$_{12}$	0.737	24.485	0.079	11.074
ARIMA(0,1,1)(1,1,0)$_{12}$	0.741	25.944	0.055	11.060

表 2　ARIMA(0,1,1)(1,1,0)$_{12}$参数估计

参数	估计值	SE	t 值	P 值
常数	0.313	5.082	0.062	0.951
差分	1	–	–	–
MA 滞后 1	0.733	0.076	9.642	<0.001
AR,季节性滞后 1	-0.344	0.104	-3.322	0.001
季节性差分	1	–	–	–

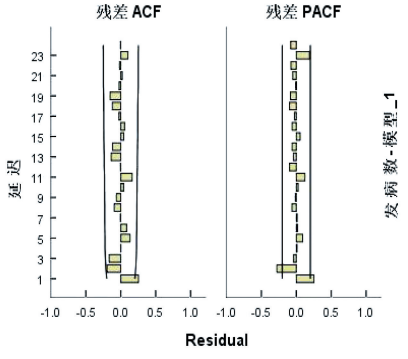


图 5　ARIMA(0,1,1)(1,1,0)$_{12}$模型残差序列的自相关和偏自相关

2.4　模型评价　模型 ARIMA(0,1,1)(1,1,0)$_{12}$整体拟合效果较好,发病趋势及季节性与实际一致(图 6),并且实际发病数据均在置信区间内。利用本模型计算甘肃省 2019 年其他感染性腹泻的发病情况并与实际发病资料进行比较评价模型,平均误差为 0.15,见表 3。

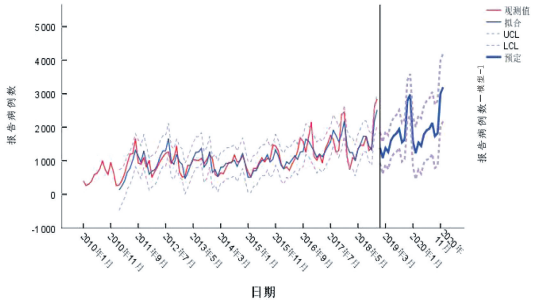


图 6　ARIMA(0,1,1)(1,1,0)$_{12}$模型拟合效果图

表 3　2019 年甘肃省其他感染性腹泻拟合发病数与实际发病数比较

项目	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
实际值	1 887	1 351	1 088	1 088	1 121	1 440	1 544	1 768	1 330	1 310	2 168	2 530
拟合值	1 392	1 084	1 382	1 256	1 609	1 744	1 817	1 945	1 549	1 634	2 802	2 973
绝对误差	-495	-267	294	168	488	304	273	177	219	324	634	443
相对误差	-0.26	-0.19	0.27	0.15	0.44	0.21	0.18	0.10	0.16	0.25	0.29	0.18

2.5　2020 年甘肃省其他感染性腹泻发病预测　预测结果显示甘肃省 2020 年其他感染性腹泻将继续增加,时间分布与往年一致,均在 8 月和 11 月左右出现发病高峰(图 6,表 4)。

表 4　2020 年甘肃省其他感染性腹泻发病数预测结果

项目	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
预测值	1 568	1 224	1 545	1 441	1 787	1 912	1 957	2 129	1 730	1 822	3 006	3 195
UCL	2 317	2 001	2 348	2 270	2 640	2 789	2 858	3 053	2 676	2 789	3 995	4 205
LCL	819	448	742	613	934	1 035	1 056	1 206	784	854	2 017	2 185

3　讨　论

其他感染性腹泻可由病毒、细菌、真菌、原虫等病原体感染,其流行面广、发病率高,严重危害人民群众健康,发病率为我国法定报告丙类传染病第 2 位,造成了严重的疾病负担,且主要集中在 5 岁以下儿童[10]。涂正波等[12]研究发现其他感染性腹泻的门诊病例和住院病例经济负担中位数分别为 1 114.5 和 3 525 元。有数据显示[4]:我国 5 岁以下儿童其他感染性腹泻发病率逐年上升。近年来甘肃省其他感染性腹泻占感染性腹泻的比例逐年增加,由 2015 年的 66%增加到 2018 年的 82%[5],因此科学地预测为其他感染性腹泻的防治提供了依据。

本研究基于甘肃省 2010—2018 年其他感染性腹泻的发病资料进行拟合分析,共收集 9 个周期数据。BIC 用来选择最优模型,一般 BIC 值越小的模型认为是最佳模型[9]。最终选定 BIC 值最小的 ARIMA(0,1,1)(1,1,0)$_{12}$模型为拟合甘肃省其他感染性腹泻发病的最优模型,其 R^2 = 0.741,Ljung-Box 检验值为 25.944。该模型拟合所得发病趋势、季节性与实际基本一致,MAPE 为 17.297%,低于王金娜等[9]利用永嘉县其他感染性腹泻数据建立 ARIMA 模型得到的 MAPE 值为 36.166%。,进一步说明此模型对实际数据的拟合准确性较高。另外,通过利用 2019 年数据验证模型,平均相对误差为 0.15,时间分布与往年一致。也有学者认为,预测的成功关键是能否判定疾病的流行程度及走向,能否为疾病的大流行和暴发提出预警[15]。

季节 ARIMA 模型结合了 ARIMA 模型和随机季节