Advances in Applied Mathematics 应用数学进展, 2021, 10(4), 1233-1242
Published Online April 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.104134

Hans汉斯

SARIMA模型在新疆布鲁氏菌病发病预测中的应用

张　隆，邢喜民，徐加波*

新疆工程学院数理学院，新疆　乌鲁木齐
Email: zl16@xjie.edu.cn, *xjb@xjie.edu.cn

收稿日期：2021年3月25日；录用日期：2021年4月15日；发布日期：2021年4月28日

摘　要

本文收集了2004年1月~2016年12月的新疆布病的月发病数，通过对数据进行序列平稳化、模型识别、模型检验的处理，预测了2017年12个月的布病发病数，拟合了2016年2月~12月的值，并与2016年2月~12月的实际值比较，最终建立了SARIMA(1,1,0)(0,1,0)12模型(AIC = 1606.44)，具有较高的有效性和合理性，该模型较好地拟合了新疆布病的新发病数，认为SARIMA模型可用于布病的短期预测和有效预防。

关键词

布鲁氏菌病，SARIMA模型，预测

Application of SARIMA Model in Prediction of Brucellosis in Xinjiang

Long Zhang, Ximin Xing, Jiabo Xu*

College of Mathematics and Physics, Xinjiang Institute of Engineering, Urumqi Xinjiang
Email: zl16@xjie.edu.cn, *xjb@xjie.edu.cn

Received: Mar. 25th, 2021; accepted: Apr. 15th, 2021; published: Apr. 28th, 2021

Abstract

In order to fit the new incidence of brucellosis in Xinjiang, this paper uses ARIMA(P,D,q)(P,D,Q)12 model to make short-term prediction and discusses the feasibility of the model. This paper collects

*通讯作者。

文章引用: 张隆, 邢喜民, 徐加波. SARIMA 模型在新疆布鲁氏菌病发病预测中的应用[J]. 应用数学进展, 2021, 10(4): 1233-1242. DOI: 10.12677/aam.2021.104134

the monthly incidence of human brucellosis in Xinjiang from January 2004 to December 2016, and uses R software to find the optimal model and make prediction. First, the incidence of brucellosis in the 12 months of 2017 is predicted. Secondly, the value from February to December 2016 is fitted, and compared with the actual value from February to December 2016, the ARIMA(1,1,0)(0,1,0)12 model (AIC = 1606.44) is finally established, which has higher effectiveness and rationality. The model fits the new incidence of human brucellosis in Xinjiang well, and can be used for short-term prediction and effective prevention of brucellosis.

Keywords

Brucellosis, SARIMA Model, Prediction

1. 引言

在过去 30 多年中，平均 2 年就有一种新的能够影响人类健康的传染病出现，在这些新发传染病中有75%都是人畜共患病[1]。布鲁氏菌病(简称布病)是由布鲁氏菌引起人和动物的一种共患性传染病，1905年在我国上海首次被发现[2]，它是由布鲁氏菌属的小型革兰氏阴性球杆菌引起的慢性、人畜共患传染病，传染源主要是病畜，感染动物可长期带菌，成为对动物和人类最危险的传染源[3]。

新疆属于农业部布病防控区域化管理的一类地区，人间布病疫情呈较高的发病态势[4]。在新疆多数人偏爱以牛羊肉为食，由于布病影响食品安全问题，所以预测布病的发病率有重要的意义。一是可以根据布病预测数据，有针对性地开展防治工作，有助于提高布病预防控制工作的能力；二是在疫情监测工作中，根据预测数据的置信区间，可以判断实际发病率是否在正常范围波动。一般年份(或月份)，传染病的发病率按照既往的变化规律如线性趋势、季节性、周期性发生变动。如果实际发病率在预测值 95%置信区间内波动，表明当月疫情基本正常，如果超出预测值 95%置信区间，表明当月疫情已不同于以往流行规律，应警惕传染病暴发或流行的可能。因此，对传染病数据作预测是非常有必要的，常用的预测模型方法有时间序列法、灰色系统法、人工神经网络法等。

本文主要是用的时间序列法，例如有一些学者用了一些方法研究传染病的流行趋势，潘姣姣等人(2012)分别用曲线回归法、指数平滑法和 ARIMA 模型模拟肺结核疫情的动态轨迹，结果显示其中 ARIMA 模型有效拟合了类似肺结核发病率的动态趋势[5]。陆波等人(2014)的研究，得出 ABIMA 模型能够有效预测流感[6]。陈纯等人(2016)用 R 软件中的 Holt-Winter 指数匀滑模型和 SARIMA 有效地预测了广州市手足口病的发病情况，结果 SARIMA 为较佳的预测分析模型[7]；易燕飞(2016)用了 ARIMA 时间序列模型和 ARIMA 乘积模型，预测了乙类传染病中的乙肝、结核病和丙类传染病中的流行性感冒的传染病流行趋势，结果显示 ARIMA 乘积模型的预测效果优于 ARIMA 模型[8]。徐秦琴等人(2017)用 SARIMA 模型较好地拟合了淄博市流行性腮腺炎的动态变化[9]。汪鹏等人(2018)比较 ARIMA 模型和 Holt-Winters 模型在武汉市流感样病例预测中的应用，结果 ARIMA 模型拟合效果较好，预测精度更高[10]。

本文选用新疆布病发病数作为研究对象，主要用的时间序列的方法对新疆布病作拟合，并预测。其中步骤为原始序列平稳性检验、模型识别、参数估计及模型诊断与优化、模型预测。

2. 资料与方法

2.1. 数据来源

如图 1 所示，新疆维吾尔自治区统计局报道的 2004 年 1 月~2016 年 12 月期间新疆人间布病的月发病数，可以看出每年发病数有上升的趋势，在 6 月份前后为高峰期。

2.2. 模型的建立

1) 季节效应分析

如图 2 所示，季节分解后，有上升的趋势性。有明显的季节性，并且以 1 年为周期，而且每年的布病高发季节和低发季节都很稳定。

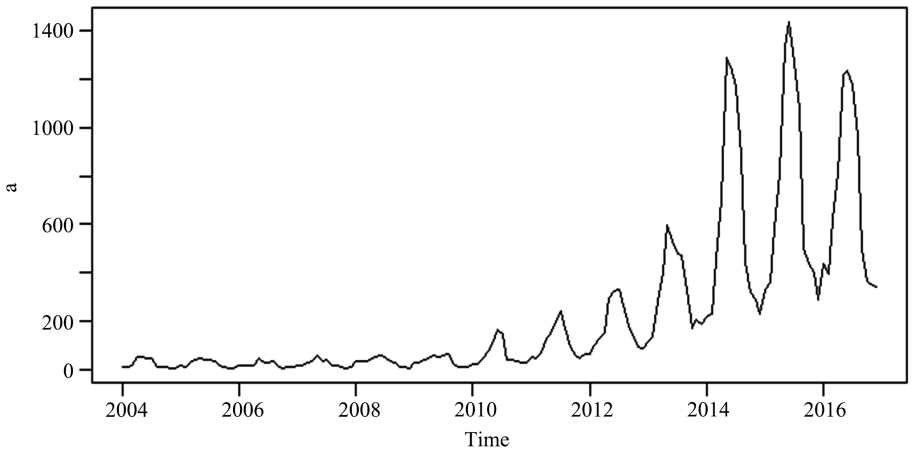

Figure 1. Monthly incidence of brucellosis in Xinjiang from January 2004 to December 2016 (a)
图 1. 2004 年 1 月~2016 年 12 月新疆布鲁氏菌病月发病数(a)

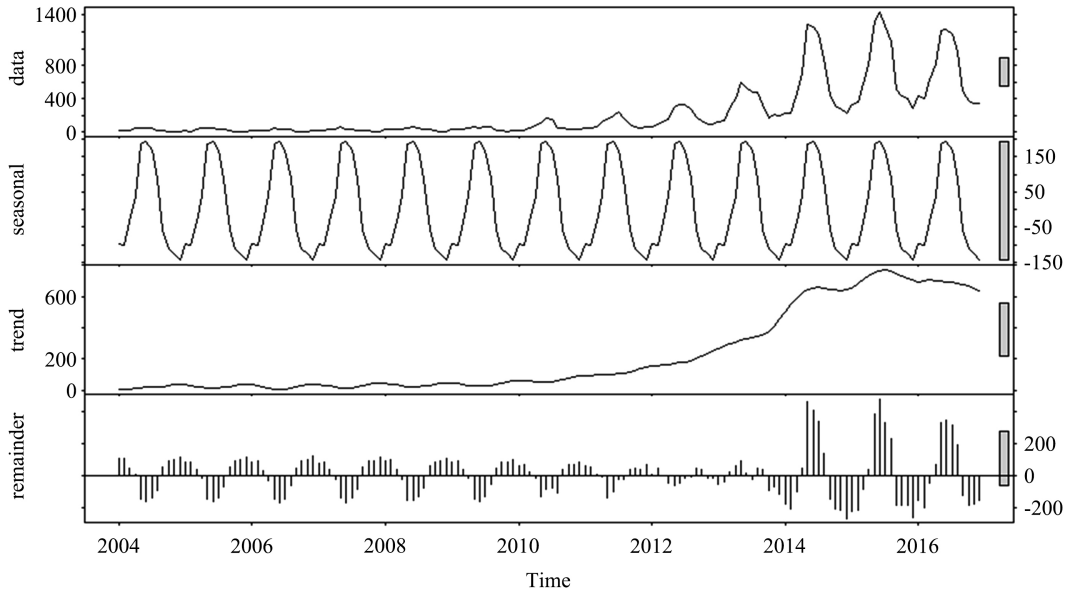

Figure 2. Decomposition of seasonal factors of brucellosis in Xinjiang
图 2. 新疆布病季节因素分解

2) SARIMA 模型介绍

SARIMA 模型：较早的文献也称其为乘积 ARIMA 模型，是随机季节模型与 ARIMA 模型的结合，对于时间序列{Z, t = 1, 2, ···}有季节性、趋势性和周期性时，可以建立非平稳季节模型，表示为 SARIMA (p,d,q)(P,D,Q)的模型，其一般形式为[11]：

$$\phi_p(L)\Phi_p\left(L^s\right)(1-L)^d\left(1-L^s\right)^D Z_t = \theta_q(L)\Theta_Q\left(L^s\right)\varepsilon_t$$

其中：

$$\phi_p(L) = 1-\phi_1 L-\phi_2 L^2-\cdots-\phi_p L^p$$

$$\Phi_P\left(L^s\right) = 1-\phi_s L^s-\phi_{2s} L^{2s}-\cdots-\phi_{Ps} L^{Ps}$$

$$\theta_q(L) = 1-\theta_1 L-\theta_2 L^2-\cdots-\theta_q L^q$$

$$\Theta_Q\left(L^s\right) = 1-\Theta_s L^s-\Theta_{2s} L^{2s}-\cdots-\Theta_{Qs} L^{Qs}$$

p 为非季节自回归阶数，P 为季节自回归阶数，q 为非季节移动平均阶数，Q 为季节自回归阶数。d,D 分别为普通差分和季节差分的阶数，s 为季节的长度，ε_t 为白噪声序列。

3) SARIMA 模型建立步骤

根据对研究序列的分析可确定和建立适当的模型：

a) 原始序列检验：ADF 单位根检验，当 P < 0.05 时可认为序列平稳。

b) 非平稳序列平稳化：根据平稳序列 acf 图和偏自相关系数 pacf 图，选择适当的阶数。

c) SARIMA 模型识别，模型识别过程中为了避免因经验不足而导致的模型识别不准确问题，使用 R 软件 auto.arima 函数自动识别模型阶数，并给出模型参数[12]。

d) 参数估计及模型诊断与优化：运用最大似然估计，充分利用序列的信息对模型中未知参数进行估计。模型检验参数的显著性检验，当 P < 0.05 时可认为参数显著。通过模型检验的 SARIMA(p,d,q)(P,D,Q)12 模型，可采用赤则准则(AIC)，贝叶斯信息准则(BIC)确定最优模型。

e) 模型预测：选择最优模型，在 80%和 95%的置信区间进行短期预测。

预测过程包括选择方法或者模型来拟合数据，然后根据拟合的模型进行预测，预测的好坏，用预测精度来度量，该文用了两种误差，如下：

均方根误差

$$\mathrm{RMSE} = \sqrt{n^{-1}\sum_{t=1}^{n}\left(e_t\right)^2}$$

平均绝对标准化误差

$$\mathrm{MASE} = n^{-1}\sum_{t=1}^{n}\left|e_t\right|/q$$

(在 MASE 中，q 对不同的对象有不同的意义，下面是针对季节性时间序列)

$$q = \frac{1}{n-m}\sum_{t=m+1}^{n}\left|x_t - x_{t-m}\right|$$

3. 结果

3.1. 原始序列平稳性检验

原始序列为非白噪声序列才有研究的意义。采用了 Ljung-Box 检验法，结果 P < 0.05，所以该序列为

非白噪声。用 ADF 单位根检验法，显示原序列为非平稳序列，所以先要非平稳序列平稳化。

如图 3 所示，分别将原序列一阶差分、一阶季节差分、一阶差分后再季节差分，结果显示一阶差分后再季节差分后的序列，经过 ADF 单位根检验，P = 0.01 说明差分后序列平稳。

3.2. 模型识别

如图 4 所示，是一阶差分在一阶季节差分后平稳状态后的 acf 图和 pacf 图，所以可确定在 SARIMA (p,d,q)(P,D,Q)12 模型中 d = 1，D = 1，看图可分析出 p 可能取值为 0 或 1，q 可能取值为 1 或 2，P 可能取值为 0 或 1，Q 可能取值为 0 或 1，也可以用 auto.arima 函数，它可以实现 ARIMA 模型的初步定阶，识别为 SARIMA(0,1,1)(0,1,0)12 为当前模型[13]。

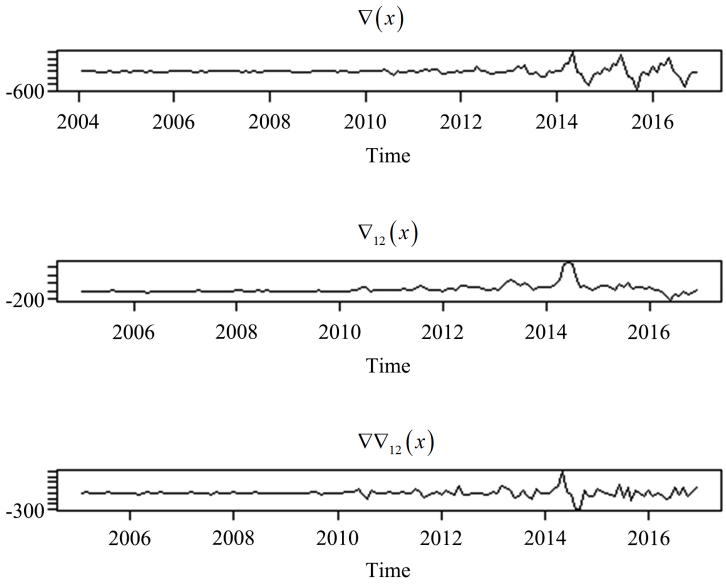

Figure 3. Various difference graphs of the sequence of a: $\{\nabla X_t\}$ (top), $\{\nabla_{12} X_t\}$ (middle), $\{\nabla\nabla_{12} X_t\}$ (bottom)

图 3. a 序列的各种差分图：$\{\nabla X_t\}$ (上)，$\{\nabla_{12} X_t\}$ (中)，$\{\nabla\nabla_{12} X_t\}$ (下)

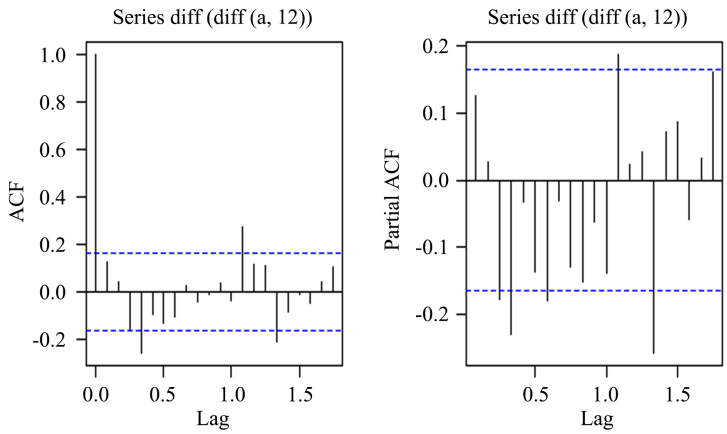

Figure 4. acf and pacf graphs after first-order difference and second-order seasonal difference

图 4. 一阶差分再一阶季节差分后的 acf 和 pacf 图

3.3. 参数估计及模型诊断与优化

3.3.1. 参数估计

SARIMA(p,d,q)(P,D,Q)12 模型可能的组合结果如表 1 所示。首先考虑建立 SARIMA(2,1,2)(1,1,1)12 模型，若显著性水平 $\alpha = 0.1$，其中变量 MA(1)的 t 值 = −0.3605、P = 0.3595 > 0.1，SAR(1)的 t 值 = 0.4837、P = 0.3147 > 0.1、SMA(1)的 t 值 = −0.5645、P = 0.2866 > 0.1，三者都没通过 t 检验。然后剔除变量 MA(1)、SAR(1)、SMA(1)，尝试建立 SARIMA(2,1,0)(0,1,0)12 模型，变量 AR(2)的 t 值 = 0.3238、P = 0.3733 > 0.1，所以剔除 AR(2)，建立 SARIMA(1,1,2)(0,1,0)12，变量 AR(1)、MA(1)不显著，剔除 AR(1)，建立 SARIMA (0,1,2)(0,1,0)12，变量 MA(2)不显著，所以剔除 MA(2)，建立 SARIMA(0,1,1)(0,1,0)12，t 检验通过，在这提一下，R 语言里的 auto.arima 函数可以帮助我们找到合适的模型，也就是它的参数检验都通过，刚好该函数选择的模型就是 SARIMA(0,1,1)(0,1,0)12，接着建立 SARIMA(1,1,0)(0,1,0) [12]模型，t 检验也通过。

Table 1. Parameter estimation and model diagnosis of SARIMA model
表 1. SARIMA 模型的参数估计和模型诊断

变量	参数检验		
	系数	t 值	P 值
SARIMA(2,1,2)(1,1,1)12			
AR(1)	0.1439	2.2449	0.0131
AR(2)	−0.7892	−13.2639	0.0000
MA(1)	−0.0115	−0.3605	0.3595
MA(2)	1.0000	29.0698	0.0000
SAR(1)	0.4371	0.4837	0.3147
SMA(1)	−0.4948	−0.5645	0.2866
SARIMA(2,1,0)(0,1,0)12			
AR(1)	0.1243	1.4798	0.0705
AR(2)	0.0271	0.3238	0.3733
SARIMA(1,1,2)(0,1,0)12			
AR(1)	−0.1652	−0.5756	0.2828
MA(1)	0.3040	1.1075	0.1349
MA(2)	0.1696	1.5224	0.0650
SARIMA(0,1,2)(0,1,0)12			
MA(1)	0.1497	1.7780	0.0387
MA(2)	0.1417	1.2743	0.1022
SARIMA(0,1,1)(0,1,0)12			
MA(1)	0.1144	1.4799	0.0705
SARIMA(1,1,0)(0,1,0)12			
AR(1)	0.1278	1.5342	0.0635

3.3.2. 模型诊断与优化

分别对SARIMA(0,1,1)(0,1,0)12模型和SARIMA(1,1,0)(0,1,0)12模型的残差进行白噪声检验，如图5、图6所示，白噪声检验主要采用了残差的密度直方图和密度估计、正态QQ图。这两个模型的直方图的残差密度有点对称，正态QQ图也类似正态分布，所以两个模型的残差都类似于白噪声。

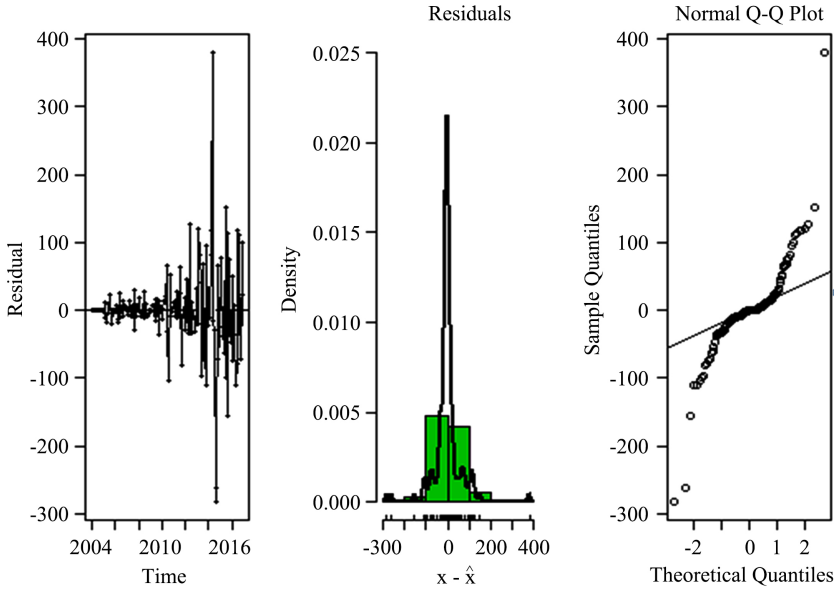

Figure 5. Residual sequence (left), density histogram and density estimation (middle), normal QQ graph (right) of SARIMA(0,1,1)(0,1,0)12 model
图5. SARIMA(0,1,1)(0,1,0)12模型的残差序列(左)、密度直方图和密度估计(中)、正态QQ图(右)

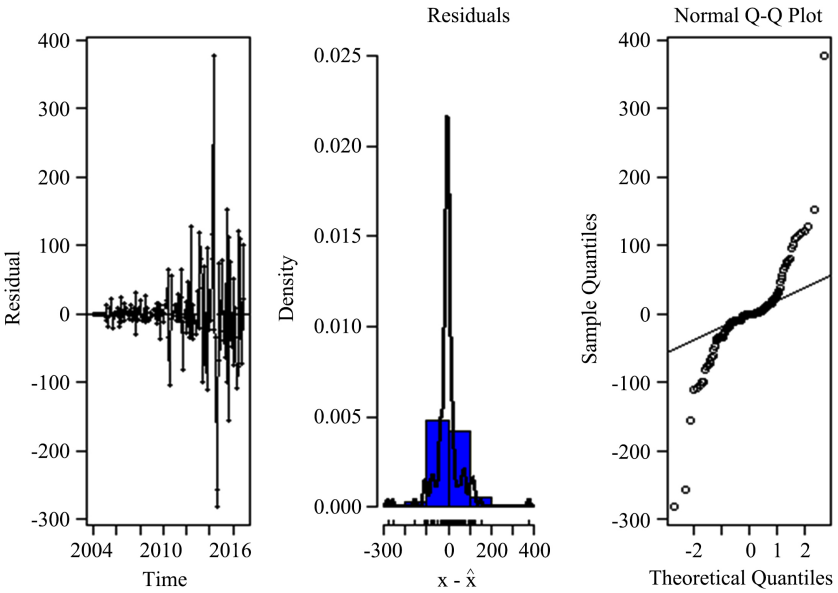

Figure 6. Residual sequence (left), density histogram and density estimation (middle), normal QQ graph (right) of SARIMA(1,1,0)(0,1,0)12 model
图6. SARIMA(1,1,0)(0,1,0)12模型的残差序列(左)、密度直方图和密度估计(中)、正态QQ图(右)

如表 2 所示，针对这两个模型可通过 AIC 和 BIC 的值越小原理 RMSE 和 MASE 的值最小选取最优模型，从而可将 SARIMA(1,1,0)(0,1,0)12 模型最初判定为最优模型 AIC = 1606.44，BIC = 1612.362。模型参数的系数也都通过检验 P < 0.1，最终 SARIMA(1,1,0)(0,1,0)12 模型被确认为最优模型。

Table 2. Selection criteria test of optimal SARIMA model
表 2. 最优 SARIMA 模型的选取准则检验

模型	AIC	BIC	RMSE	MASE
SARIMA(0,1,1)(0,1,0)12	1606.66	1612.583	62.8786	0.576
SARIMA(1,1,0)(0,1,0)12	1606.44	1612.362	62.8294	0.574

3.4. 模型预测

1) 预测 2017 年 1 月~2017 年 12 月的新发病数

如图 7 所示，采用 SARIMA(1,1,0)(0,1,0)12 模型，预测新疆人间布病 2017 年 1 月~2017 年 12 月新发病数。如表 3 所示，2017 年 1 月~2017 年 11 月的预测值以及预测区间。

2) 2016 年 1 月以前的数据作为训练集，2016 年 2 月以后的数据作为测试集

如图 8 所示，采用 SARIMA(1,1,0)(0,1,0)12 模型，通过 2004 年 1 月~2016 年 1 月的数据(训练集)拟合并预测新疆人间布病 2016 年 2 月~2016 年 12 月新发病数(测试集)。如表 4 所示，2016 年 2 月-2016 年 12 月的预测值以及预测区间。如表 5 所示，训练集与测试集的模型交叉验证，可以看出该模型较为合理的拟合出新疆新发的人间布病数量。

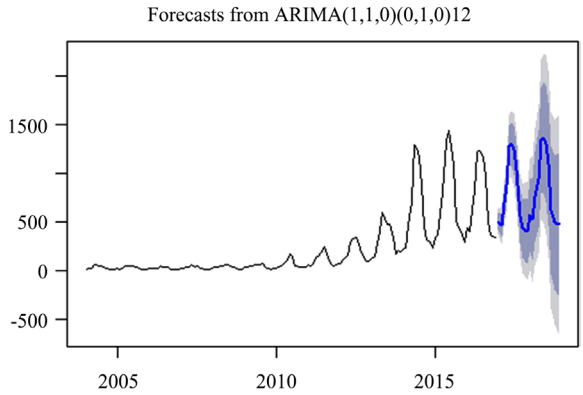

Figure 7. The ARIMA(1,1,0)(0,1,0)12 model is proposed to predict the new incidence of human brucellosis in Xinjiang from January to December 2017
图 7. SARIMA(1,1,0)(0,1,0)12 模型拟合并预测新疆人间布病 2017 年 1 月-12 月新发病数

Table 3. The ARIMA(1,1,0)(0,1,0)12 model with 80% and 95% confidence intervals predicts the number of new cases of human brucellosis in Xinjiang from January 2017 to November 2017
表 3. 在置信区间 80%和 95%下的 SARIMA(1,1,0)(0,1,0)12 模型预测新疆人间布病 2017 年 1 月~2017 年 11 月新发病数

	17-1	17-2	17-3	17-4	17-5	17-6	17-7	17-8	17-9	17-10	17-11
预测值	501	460	686	874	1280	1297	1238	1028	558	431	411
80%上	417	333	527	688	1071	1066	988	760	273	130	95
80%下	585	587	845	1060	1490	1528	1488	1296	843	732	727
95%上	373	266	443	590	960	944	856	618	123	−29	−72
95%下	630	654	929	1159	1601	1650	1620	1438	994	891	894

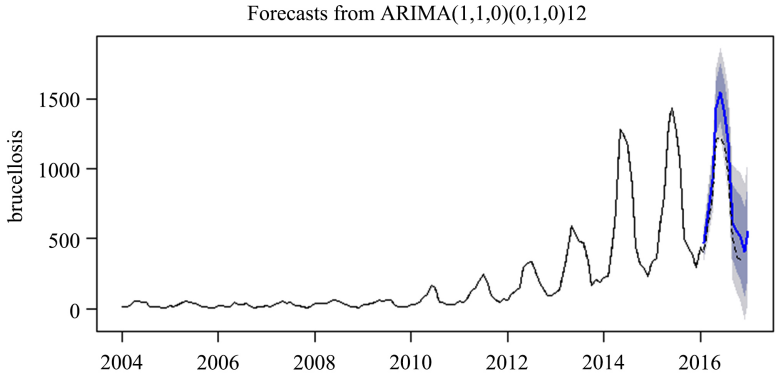

Figure 8. The ARIMA(1,1,0)(0,1,0)12 model is proposed to predict the new incidence of human brucellosis in Xinjiang from February to December 2016, and compare with the actual value from February to December 2016
图 8. SARIMA(1,1,0)(0,1,0)12 模型拟合并预测新疆人间布病 2016 年 2 月~12 月新发病数，以及与 2016 年 2 月~12 月实际值比较

Table 4. SARIMA(1,1,0)(0,1,0)12 model predicted and actual values under confidence intervals of 80% and 95%
表 4. 在置信区间 80%和 95%下的 SARIMA(1,1,0)(0,1,0)12 模型预测值和实际值

	2-16	3-16	4-16	5-16	6-16	7-16	8-16	9-16	10-16	11-16	12-16
预测值	471	699	922	1442	1551	1383	1194	618	550	516	406
实际值	396	622	810	1216	1233	1174	964	494	367	347	340
80%上	388	573	762	1255	1340	1150	941	347	262	211	86
80%下	554	826	1082	1630	1763	1617	1447	890	839	821	727
95%上	344	506	678	1155	1228	1026	807	203	109	50	−84
95%下	598	893	1167	1729	1875	1740	1581	1034	992	983	896

Table 5. Model cross validation
表 5. 模型交叉验证

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	6.70	31.90	17.71	−5.66	24.52	0.47	0.33	NA
Test set	263.17	320.72	265.29	35.83	36.75	7.06	0.67	1.03

4. 讨论

本文研究表明，在新疆布病研究中，SARIMA 模型的建立过程可以分为四步：原始数据平稳化检验；模型识别；参数估计及模型诊断与优化；模型预测。我们使用 SARIMA(p,d,q)(P,D,Q)12 模型来分析新疆布病数据。最后用 SARIMA(1,1,0)(0,1,0)12 最优模型来预测布病新发病数。结果显示预测值与实际值都在 80%置信区间和 95%置信区间波动，这表明模型是合理的，且预测值是有效的。对 SARIMA(1,1,0)(0,1,0)12 模型的残差序列进行白噪声检验，判断该模型的合适性。其残差序列的检验值远大于判定值 0.05，说明该模型的残差序列为白噪声序列，残差序列中有用的信息已被提取完，模型拟合程度很好。这证实了 SARIMA 模型的可行性。SARIMA 模型考虑了时间序列的周期性和季度性变化，有较高的预测精度。但由于数据规模的限制，随着预测时间的延长，预测误差会逐渐增大，预测精度也会下降。为了比较 SARIMA

(0,1,1)(0,1,0)12 模型和 SARIMA(1,1,0)(0,1,0)12 模型的预测精度，我们分析了两个模型的残差。因此，通过 RMSE，MASE，AIC，BIC 标准测试出最好的模型是 SARIMA(1,1,0)(0,1,0)12 模型。

利用 SARIMA 模型对疾病进行预测分析，有两大优点：1) 可以不考虑其他相关因素，只用考虑变量的自生变化；2) 可以用 R 语言的 auto.arima 函数自动定阶，出来个合适的模型，然后再去找最优模型。但是我们还应该注意：1) 至少需要 50 个以上的历史数据；2) 建立的模型只适合当前数据的短期预测，模型要随时更新。对于已建立的模型应不断加入新的实际值，以修正或重新拟合更优的模型[14]。

学者们用不同模型来拟合一些传染病数据，并作预测，大部分为了比较这些方法的原理和实践中的差异。研究结果表明，尽管没有发现任何一种方法比其他方法更好，在未来的工作中，我们可以考虑其他模型应用于布病的研究，并寻求更精确的模型来预测新疆布病的发病率。

基金项目

新疆维吾尔自治区高校科研计划自然科学项目(XJEDU2021Y048)，新疆工程学院博士启动基金(2020xgy012302)共同资助。

参考文献

[1] 张艳红. 人畜共患病的流行特点[J]. 畜禽业, 2013(6): 8-9.

[2] Shang, D.Q., Xiao, D.L. and Yin, J.M. (2002) Epidemiology and Control of Brucellosis in China. *Veterinary Microbiology*, **90**, 165-182. https://doi.org/10.1016/S0378-1135(02)00252-3

[3] 陈彪, 王涛, 李爱巧, 等. 乌鲁木齐市动物布鲁氏菌病流行病学调查[J]. 中国动物检疫, 2013, 30(3): 28-30.

[4] 木合塔尔·艾山, 何海波, 邰新平, 等. 新疆 2013 年人间布鲁氏菌病监测结果及疫情分析[J]. 中国媒介生物学及控制杂志, 2015, 26(1): 86-88.

[5] 潘姣姣, 董柏青, 吕炜, 等. 三种时间序列模型探讨 1989~2012 广西肺结核发病趋势[J]. 中国卫生统计, 2012, 29(6): 868-870.

[6] 陆波, 闵红星, 扈学琴, 等. 时间序列模型预测流感发病率的研究[J]. 中国实用医药, 2014, 9(7): 255-256.

[7] 陈纯, 李铁钢, 肖新才, 等. 应用 R 软件对比两种手足口病发病预测模型的效果[J]. 国际流行病学传染病学杂志, 2016, 43(2): 101-104.

[8] 易燕飞. 基于时间序列模型的传染病流行趋势及预测研究[D]: [硕士学位论文]. 长春: 长春工业大学, 2016.

[9] Xu, Q.Q., Li, R.Z., Liu, Y.F., Luo, C., Xu, A.Q., Xue, F.Z., Xu, Q. and Li, X.J. (2017) Forecasting the Incidence of Mumps in Zibo City Based on a SARIMA Model. *International Journal of Environmental Research and Public Health*, **14**, 925. https://doi.org/10.3390/ijerph14080925

[10] 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.

[11] 吴喜之, 刘苗. 应用时间序列分析[M]. 第 2 版. 北京: 机械工业出版社, 2018: 38-39.

[12] 妥小青, 张占林, 龚政, 等. 基于 ARIMAX 模型的乌鲁木齐市流感样病例预测分析[J]. 中华疾病控制杂志, 2018, 22(6): 590-593.

[13] Hyndman, R.J. and Khandakar, Y. (2008) Automatic Time Series Forecasting: the forecast Package for R. *Statistical Software*, **27**, 16. https://doi.org/10.18637/jss.v027.i03

[14] 漆莉, 李革, 李勤. ARIMA 模型在流行性感冒预测中的应用[J]. 第三军医大学学报, 2007, 29(3): 267-269.