# 基于 ARIMA 模型的江苏省梅毒疫情预测

张文娟[1],刘文东[2],胡建利[2],汤奋扬[2],彭志行[1],喻荣彬[1*]

([1]南京医科大学公共卫生学院流行病与卫生统计学系,江苏 南京 211166;[2]江苏省疾病预防控制中心,江苏 南京 210009)

[摘 要] **目的**:了解江苏省梅毒的流行病学特点,构建预测江苏省梅毒月发病率的自回归移动平均模型(autoregressive integrated moving average model,ARIMA),为梅毒预防控制工作提供参考依据。**方法**:利用江苏省 1995—2009 年梅毒月发病率资料建立 ARIMA 预测模型,并进行模型评价。**结果**:拟合 ARIMA(1,1,0),(2,1,0)模型为预测江苏省梅毒月发病率的最佳模型,模型周期性、季节性一阶、二阶系数分别为–0.579、–0.245、–0.357,$t$ 检验统计量分别为 8.777,2.881,4.766,相应的 $P$ 值分别为<0.001、0.005、<0.001,表明该模型具有较高的预测精度,预测值与实际值较为接近,且实际值均在预测值的 95%置信区间范围内,预测效果较好。**结论**:ARIMA 模型能较好地预测梅毒发病率的变化趋势,为梅毒预防控制措施的制定提供重要依据。

[关键词] ARIMA 模型;时间序列分析;梅毒;预测

[中图分类号] R759.1 [文献标志码] A [文章编号] 1007-4368(2017)05-0649-04

**doi**:10.7655/NYDXBNS20170531

## Forecast of syphilis epidemic situation in Jiangsu Province base on ARIMA model

Zhang Wenjuan[1],Liu Wendong[2],Hu Jianli[2],Tang Fenyang[2],Peng Zhihang[1],Yu Rongbin[1*]

*([1]Department of Epidemiology and Biostatistics,School of Public Health,NJMU,Nanjing 211166;[2]Jiangsu Provincial Center for Disease Control and Prevention,Nanjing 210009,China)*

[**Abstract**] **Objective**:To investigate the epidemiologic characteristics of syphilis,establish an Autoregressive Integrated Moving Average Model (ARIMA) model for the prediction of monthly incidence of syphilis in Jiangsu and provide evidence for the prevention and control of the disease. **Methods**:We used monthly incidence data of syphilis in Jiangsu from 1995 to 2009 to establish the ARIMA model,and then evaluated the model. **Results**:ARIMA (1,1,0),(2,1,0) models are the optimal models to predict the monthly incidence of syphilis in Jiangsu,the coefficients of recurrent model,seasonal first-order model,seasonal second-order model respectively are –0.579,–0.245,–0.357,statistics of t test respectively are 8.777,2.881,4.766. Correspondingly,the values of $P$ respectively are 0.000,0.005,<0.001. The model had favorably high precision,the predicting value was close to the true value,which was with in the 95% confidence interval of the predicting value. **Conclusion**:The ARIMA model was suitable to forecast the incidence of syphilis. ARIMA model could be used to predict the incidence trend of syphilis and provide evidence for the development of syphilis prevention and control measures.

[**Key words**] ARIMA model;time series analysis;syphilis;predict

[Acta Univ Med Nanjing,2017,37(05):649-652]

梅毒(syphilis)是由梅毒螺旋体引起的性传播疾病,是性病中危害较严重的一种,被我国列为法定乙类传染病。近年来,我国的梅毒发病率迅速上升,且增长速度逐渐加快。江苏省地处我国东部,梅毒发病数从 2004 年全省上报的 7 692 人次增长到 2009 年的 23 977 人次,年平均增长 25.53%,报告病例数在全国一直排第 4~5 位[1,2]。2004—2008 年,报告病例中各期梅毒均逐年增加,以早期显性(一、二期)梅毒为主,占 69.75%,发病年龄集中在 20~39 岁年龄组,占 51.55%,在经济较发达的苏南地区,梅毒疫情更为严重,占全省报告病例数的 57.19%[2]。本研究探索梅毒疫情预测的时间序列分析方法,研究梅毒疫情的长期走势,对江苏省梅毒的月发病率进行建

[基金项目] 国家自然科学基金(81673275,U1503123);国家“十二五”重大科技专项(2012ZX10001-001);江苏省高校优势学科建设工程资助

*通信作者(Corresponding author),E-mail:rongbinyu@njmu.edu.cn

模分析，以预测梅毒未来的疫情特点、变化规律，为梅毒的预防控制提供理论依据。

## 1 资料和方法

### 1.1 资料

江苏省梅毒分月发病资料来自江苏省疾病控制网络直报系统，江苏省人口数据来自江苏省统计年鉴。

### 1.2 方法

#### 1.2.1 模型的基本思想

自回归移动平均(autoregressive integrated moving average，ARIMA) 模型用相应的数学模型描述一组依赖时间的随机变量相互之间的自相关性，以表征预测对象发展的延续性并从时序的过去值与现在值预测其未来值[3]。

#### 1.2.2 建模过程

建立 ARIMA 模型可归纳为 3 个具体步骤：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[4]。

数据的预处理：如果一个数据序列的平均值和方差始终为常数，则称它是平稳的。平稳化就是将数据图上呈线性或非线性的数据经过处理转化为平稳的数据，处理后的数据是否平稳可以用自相关函数进行判定。通常用下列方法：如果序列呈线性趋势，均值不平稳，则利用一阶差分；如果序列呈现二次趋势，均值不是常数，则利用二阶差分；如果序列呈现随时间的上升或下降偏差，方差不是常数，则通常可利用自然对数将其平稳化。

模型的识别、定阶及参数估计：利用自相关分析和偏自相关分析等方法，根据 Box-Jenkins 方法同时参考赤池信息准则(Akaike information criterion，AIC)与 Schwarz 贝叶斯准则 Schwarz Bayesian criterion，SBC)[5]的结果来确定模型的阶数，同时用最大似然估计法来估计模型参数。

模型的诊断检验：对所建模型的拟合优度进行检验以验证其是否合适，方法是对观测值和模型拟合值进行残差分析。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。此时可对残差拟合更复杂的模型以充分提炼资料的信息，从而得到更合适的模型，如果残差序列不是白噪声序列则需重新建立模型，重复上述步骤，直到残差序列是白噪声序列为止[6]。

## 2 结 果

### 2.1 江苏省梅毒疫情分析

2004—2010 年江苏省累计报告病例数最多的地级市依次为苏州市、南京市、常州市和无锡市，占全省报告病例数的 57.19%。在苏北地区，如连云港、淮安市和泰州市的报告病例数较少，但年增长率大幅度上升，年均发病率的增长率均超过 30%。梅毒发病年龄集中在 20~39 岁年龄段(占 51.55%)，其中 20~29 岁年龄段占 23.96%，30~39 岁年龄段占 27.59%。其次为 50 岁以上年龄段(占 24.82%)，接下来是 40~49 岁年龄段(占 19.60%)、14 岁以下年龄段(占 2.55%)、15~19 岁年龄段(占 1.48%)。

梅毒发病人数的时间序列分析：1995—2009 年江苏省梅毒发病率总体呈波动上升的趋势，其中在 1995—2000 年间梅毒发病率缓慢上升，在 2001—2003 年发病率略有下降，但在 2004 年之后梅毒的发病率迅速升高。另外，1995—2009 年梅毒的发病率时序图呈非线性，并且存在明显的周期性波动趋势，以 12 个月为 1 个周期，发病高峰主要集中在每年的第三季度，其次是第二季度，数据为不平稳序列。

### 2.2 建立 ARIMA 模型

根据梅毒发病率的年周期性，对 1995—2009 年江苏省梅毒的发病数据自然对数后一阶差分，季节一阶差分后的序列图可见原始数据经上述作用后数据平稳(图 1A)。

观察对数变换一阶差分后序列的自相关图、偏自相关图 (图 1B、C)，发现自相关系数在 $P$ 分别为 1，2，8 处较大，偏自相关系数在 $Q$=1，$P$=1 处较大，经试算比较后取 $Q$=1，$P$=0，$P$=2 模型的各项指标相对较好。拟合 ARIMA(1，1，0)，(2，1，0)模型为江苏省梅毒月发病率预测的最佳模型。

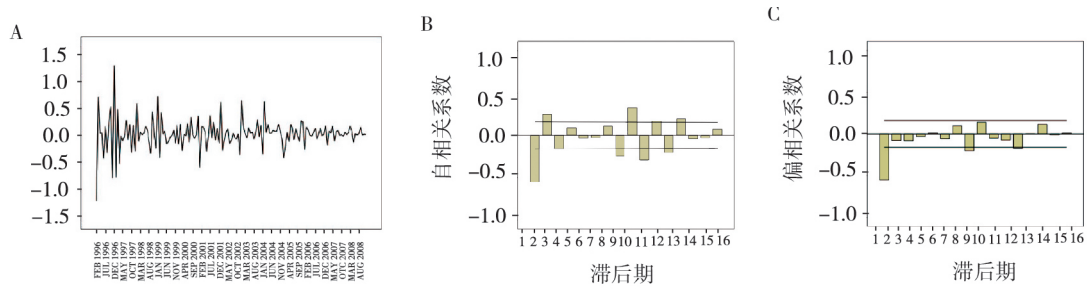
经检验，此模型Ljung-Box 统计量为 32.145，$P$=0.413，远大于 0.05，表明不拒绝原假设，模型的拟合程度很好。模型的参数估计结果显示，模型周期性、季节性一阶、二阶系数分别为 −0.579、−0.245、−0.357，$t$ 检验统计量分别为 8.777、2.881、4.766，相应的 $P$ 值为<0.001，0.005，<0.001，均<0.01，表明在 0.01 的水平下，拒绝原假设，系数显著不为 0。模型表达式为：$(1+0.579B)(1+0.245B^{12}+0.357B^{24})(1-B)(1-B^{12})\ln x_t=\varepsilon t$，模型的诊断检验，可看出残差的自相关系数和偏自相关系数基本均在置信区间内，残差为白噪声。

用此模型对 1995—2009 年江苏省梅毒数据序列进行模型数据拟合,并对 2010 年 1—12 月的发病人数进行预测,模型拟合及预测情况见图 2。可见拟合值与预测值基本落在真实值的可信区间内,表明模型选择是合适的,拟合效果较好。对模型预测的准确性检验结果见表 1。

## 3 讨 论

传统的时间序列分析模型要求序列具有平稳的线性趋势[4],江苏省梅毒的发病率存在季节性和周期性,为非线性序列数据,如果不考虑这些因素的影响,建立的统计模型就无法外推,预测结果往往不准确,没有推广应用的价值。本研究采用的 ARIMA 模型考虑了这些因素对序列平稳性造成的影响,更加针对传染病的实际发病情况,因而改善了预测结果,显示出 ARIMA 模型对非线性数据的适用性。另外,ARIMA 模型将各种复杂因素的综合效应统一蕴涵于时间变量之中,这也是时间序列分析应用于疾病预测的一个突出优点[7]。本文的预测结果可以表明,ARIMA 模型用于传染病的短期预测时,具有实用性强、精确度高的特点[8-9]。但由于传染病发病情况受很多因素影响,比如气候、人为的预防控制等,因此利用此模型对梅毒未来发病趋势的预测仅是基于以往综合因素的基础上进行的理论上的预测,具有一定的局限性。如果将一些已知的影响因素纳入模型进行分析,对模型进行适当的调整,可进一步提高模型的精确度。

1995—2010 年江苏省传染病网络直报数据显示,近年来江苏省梅毒的发病率呈快速增长趋势,发病者主要集中在青壮年年龄组,发病例数占江苏省




A:梅毒人数自然对数后一阶差分,季节一阶差分后的序列图;B:序列自相关图;C:序列偏相关图.

**图 1 江苏省梅毒发病率预测的 ARIMA 模型**

**Figure 1 ARIMA model for the prediction of incidence trend of syphilis in Jiangsu Province**

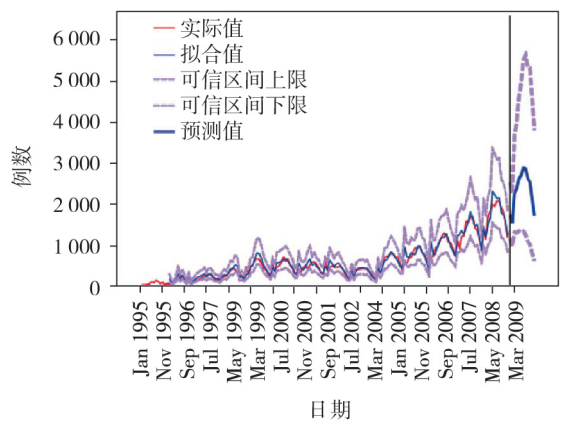



**图 2 1995—2009 年实际值、拟合值、预测值、可信区间图**

**Figure 2 The actual values,fitted values and predictive values and confidence interval graph from 1995 to 2009**

**表 1 2009 年江苏省梅毒发病率预测值与真实值比较**

**Table 1 Comparison of predicted values and true values of syphilis incidence in Jiangsu Province in 2009**

| 时间 | 真实值 | 预测值 | 相对误差 |
|---|---|---|---|
| 2009 年 1 月 | 1 445 | 1 750 | 0.211 073 |
| 2009 年 2 月 | 1 660 | 1 539 | 0.072 890 |
| 2009 年 3 月 | 1 837 | 2 282 | 0.242 243 |
| 2009 年 4 月 | 2 002 | 2 386 | 0.191 808 |
| 2009 年 5 月 | 2 218 | 2 624 | 0.183 048 |
| 2009 年 6 月 | 2 544 | 2 708 | 0.064 465 |
| 2009 年 7 月 | 2 512 | 2 905 | 0.156 449 |
| 2009 年 8 月 | 2 548 | 2 869 | 0.125 981 |
| 2009 年 9 月 | 2 103 | 2 632 | 0.251 545 |
| 2009 年 10 月 | 1 862 | 2 544 | 0.366 273 |
| 2009 年 11 月 | 1 548 | 2 162 | 0.396 641 |
| 2009 年 12 月 | 1 698 | 1 725 | 0.015 901 |

的 51.55%[4],以经济发达地区为多,经济不发达地区快速增长,提示江苏省梅毒疫情有较大范围流行的趋势,梅毒的防治工作十分严峻。本文用 ARIMA 模型对江苏省梅毒发病率的未来趋势、波动特点进行了较好的预测分析,为梅毒的预防控制工作提供了理论依据。根据本文的研究结果,参照江苏省梅毒的流行病学特点, 制定相应的预防控制措施[11]。防制时间上重点应放在第二、三季度,人群上应集中在青壮年人群,对于经济发达地区加强控制措施,对于经济相对落后地区应严防梅毒较大范围的流行,具体的措施包括宣传教育,消灭嫖娼卖淫犯罪活动等[11-12]。

**[参考文献]**


[1] 羊海涛,傅更锋,徐金水,等. 江苏省 2005—2007 年梅毒疫情核查结果分析 [J]. 中国公共卫生,2009,25(9):1133-1134

[2] 傅更锋,还锡萍,丁 萍,等. 江苏省 2004—2008 年梅毒流行病学分析及防治策略研究[J]. 南京医科大学学报(自然科学版),2009,29(10):1399-1402

[3] 彭志行,陶 红,贾成梅,等. 时间序列分析在麻疹疫情预测预警中的应用研究[J]. 中国卫生统计,2010,27(5):459-463

[4] 布洛克威尔著,田铮译. 时间序列的理论与方法[M]. 2 版. 北京:高等教育出版社,2001:13-23

[5] 郭 璐,张 敏,朱正平,等. ARIMA 模型在南京市梅毒预测中的应用[J]. 现代预防医学,2015,42(2):205-207

[6] Zhang X,Zhang T,Pei J,et al. Time series modelling of syphilis incidence in China from 2005 to 2012[J]. PLoS One,2016,11(2):e0149401

[7] Wang Y,Li X,Chai F,et al. Application of ARIMA model in prediction of incidence of syphilis in China [J]. Mod Prevent Med,2015,42(3):385-388

[8] 王永斌,李向文,柴 峰,等. ARIMA 模型在我国梅毒发病率预测中的应用[J]. 现代预防医学,2015,42(3):385-388

[9] Andersson J,Karlis D. Treating missing values in INAR(1)models:An application to syndromic surveillance data [J]. J Time Ser Anal,2010,31(1):12-19

[10] 王 娜,张馨月,张吴琼,等. 神经梅毒的诊断与治疗新进展[J]. 中国现代神经疾病杂志,2016,16(7):397-403

[11] 李 珏. 医疗机构梅毒报告病例准确性现场核查结果分析[J]. 中国艾滋病性病,2014,20(3):216-217

[12] 陈 勇,张 玲,詹永婧,等. 神经梅毒强化驱梅治疗疗效预测因素的回顾性研究[J]. 中华实验和临床感染病杂志(电子版),2016,10(3):274-279


[收稿日期] 2016-11-13

(上接第 641 页)


tients on peritoneal dialysis[J]. Kidney Res Clin Pract, 2016,35:169-175

[10] Krishnamoorthy V, Sunder S, Mahapatra H, et al. Evaluation of protein-energy wasting and inflammation on patients undergoing continuous ambulatory peritoneal Dialysis and its correlations[J]. Nephro-urology Monthly, 2015,7(6):1-8

[11] Milan Manani S, Virzì GM, Clementi A, et al. Pro-in-flammatory cytokines: a possible relationship with dialytic adequacy and serum albumin in peritoneal dialysis patients[J]. Clinical Kidney Journal,2016,9(1):153-157

[12] 周长菊,曹 娟,章 旭,等. 维持性透析患者的蛋白能量消耗情况及影响因素分析[J]. 中国血液净化,2016,15(9):483-487

[13] 孙亦兵,刘月英,王 莹. 残余肾功能对腹膜透析患者营养状况的影响[J]. 蚌埠医学院学报,2014,39(8):1078-1079

[14] John MM, Gupta A, Sharma RK, et al. Impact of residual renal function on clinical outcome and quality of life in patients on peritoneal dialysis[J]. Saudi J Kidney Dis Transpl,2017,28(1):30-35

[15] 汪 涛. 腹膜透析患者与残余肾功能[J]. 肾脏病与透析肾移植杂志,2011,20(3):256-257

[16] Sikorska D, Pawlaczyk K, Gawlik AO, et al. The importance of residual renal function in peritoneal dialysis[J]. Int Urol Nephrol,2016,48:2101-2108

[17] Munguia-Miranda C, Ventura-Garcia Mde J, Avila-Diaz M, et al. Factors related to residual renal function loss in patients in peritoneal dialysis[J]. Rev Med Inst Mex Seguro Soc,2015,53(5):578-583

[18] Ryckelynck JP, Goffin E, Verger C et al. Maintaining residual renal function in patients on dialysis [J]. Nephrol Ther,2013,9(6):403-407


[收稿日期] 2016-09-17